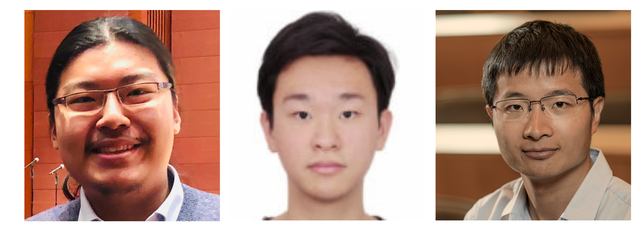
Imitating Deep Learning Dynamics via Locally Elastic Stochastic Differential Equations

Local Elasticity: training on a sample has larger effects on samples similar to it than those dissimilar to it. For example, training on an image of cat has a greater effect on other images of cats than images of, say, dogs.



Jiayao Zhang 张家耀 Hua Wang 王 华 Weijie J. Su 苏炜杰
University of Pennsylvania
{zjiayao,wanghua,suw}@wharton.upenn.edu

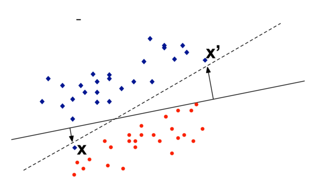
NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

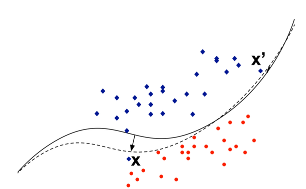
Understanding the training dynamics of deep learning models is perhaps a necessary step toward demystifying their effectiveness. In particular,

*how do data from different classes gradually become **separable in their feature spaces** when training neural networks using stochastic gradient descent?*

We take a *phenomenological approach* to model feature evolutions of neural net training using a set of stochastic differential equations (SDEs) that *each corresponds to a training sample*. Concretely, for binary classification, with m being time, superscripts class indices, subscripts sample indices, and (α, β) parameters measuring the strengths of *local elasticity*[1-3], we model



$$\begin{cases} X_i^1(m) = X_i^1(m-1) + h \cdot \alpha X_l^1(m-1) + \text{noise}, \\ X_j^2(m) = X_j^2(m-1) + h \cdot \beta X_l^1(m-1) + \text{noise}. \end{cases}$$



Our main finding uncovers a **sharp phase transition** regarding the ***intra-class*** impact: if and only if the SDEs are *locally elastic* in the sense that the impact is more significant on samples from the same class as the input, the features of the training data are asymptotically linearly separable.

The LE-SDE/LE-ODE Model

We define the *generic* feature vector $\tilde{\boldsymbol{X}}(t) = (\tilde{X}^k(t))_{k=1}^K \in \mathbb{R}^{Kp}$ as the concatenation of p-dimensional feature vectors from K classes, and model its dynamics as $\mathrm{d}\tilde{\boldsymbol{X}}(t) = \boldsymbol{M}(t)\tilde{\boldsymbol{X}}(t)\,\mathrm{d}t + \Sigma(t)\,\mathrm{d}\boldsymbol{B}_t$, where the drift $\boldsymbol{M}(t) = (\boldsymbol{E}(t) \otimes \boldsymbol{P}) \circ \boldsymbol{H}$ consists of the LE matrix $\boldsymbol{E}(t)$ that encodes the strengths of local elasticity (analogous to the magnitudes of α and β), the similarity matrix H that encodes the direction in which features interacts (analogous to the phase of α and β), and a sampling matrix P modeling randomnesses from such as mini-batch sampling and label corruption, and imbalanced datasets. Here, B_t denotes standard Brownian motion.

The simplest LE matrix consists of two values $\alpha(t)$ and $\beta(t)$. In this case, in the large sample limit, we obtain the following LE-ODE for the mean features, where $\circ$ denotes the Hadamard product (with a slight abuse of notation):

$$\mathrm{d}\bar{\boldsymbol{X}}(t) = \boldsymbol{M}(t)\bar{\boldsymbol{X}}(t)\,\mathrm{d}t = ((\boldsymbol{E}(t) \otimes \boldsymbol{P}) \circ \boldsymbol{H})\,\bar{\boldsymbol{X}}(t)\,\mathrm{d}t.$$

When $\boldsymbol{P} = \boldsymbol{1}\boldsymbol{1}^\top/K$, we write the LE-ODE as the follows where $\bar{\boldsymbol{X}} = \mathbb{E}_{\textbf{data}}\tilde{\boldsymbol{X}}$:

$$\mathrm{d}\underbrace{\begin{bmatrix} \begin{bmatrix} \bar{X}_1^1 \\ \vdots \\ \bar{X}_p^1 \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \bar{X}_1^K \\ \vdots \\ \bar{X}_p^K \end{bmatrix} \end{bmatrix}}_{\bar{\boldsymbol{X}}(t)} = \frac{1}{K} \underbrace{\begin{bmatrix} \alpha(t) & \beta(t) & \dots & \beta(t) \\ \beta(t) & \alpha(t) & \dots & \beta(t) \\ \vdots & \vdots & \dots & \vdots \\ \beta(t) & \dots & \dots & \alpha(t) \end{bmatrix}}_{\boldsymbol{E}(t)} \circ \underbrace{\begin{bmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} & \dots & \boldsymbol{H}_{1K} \\ \boldsymbol{H}_{21} & \boldsymbol{H}_{22} & \dots & \boldsymbol{H}_{2K} \\ \vdots & \dots & \dots & \vdots \\ \boldsymbol{H}_{K1} & \boldsymbol{H}_{K2} & \dots & \boldsymbol{H}_{KK} \end{bmatrix}}_{\boldsymbol{H}} \begin{bmatrix} \begin{bmatrix} \bar{X}_1^1 \\ \vdots \\ \bar{X}_p^1 \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \bar{X}_1^K \\ \vdots \\ \bar{X}_p^K \end{bmatrix} \end{bmatrix} \mathrm{d}t.$$

The Separation Theorem

Our main contributions include the following separation theorem.

Theorem (Separation of LE-SDE)

Suppose $\gamma(t) = \alpha(t) - \beta(t) > 0$, assume $\boldsymbol{H} = (\boldsymbol{H}_{ij})_{ij}$ is positive semi-definite (PSD) with positive diagonal entries. As $t \to \infty$, we have

1. *if $\gamma(t) = \omega(1/t)$, the features are separable with probability tending to 1;*
2. *if $\gamma(t) = o(1/t)$, and the number of per-class-feature n tending to ∞ at an arbitrarily slow rate, the features are asymptotically pairwise separable with probability 0.*

Here, $\gamma(t) = \omega(1/t)$ stands for $\gamma(t) \gg 1/t$ as $t \to \infty$. For example, $1/t^{0.5} = \omega(1/t)$ and $(t \ln t)^{-1} = o(1/t)$ as $t \to \infty$.

Visualizing Phase Transition

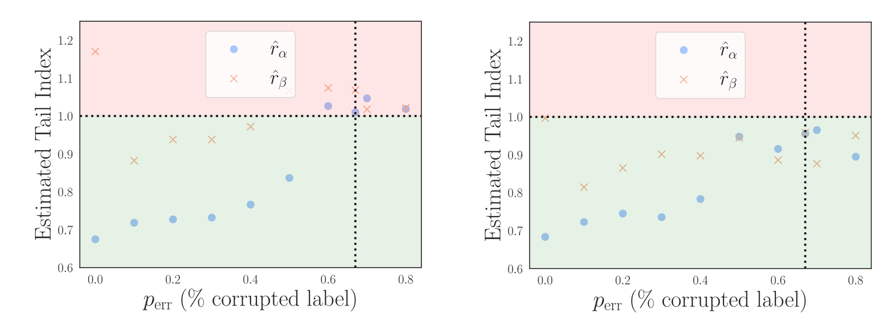


Figure 1(a): Estimated tail indices under I-model. Figure 1(b): Estimated tail indices under L-model.

Model	$\boldsymbol{H}_{ij}$	Remark
I-model	$\boldsymbol{I}_p$	Isotropic Feature Model
L-model	$\bar{\boldsymbol{H}}^j = \boldsymbol{d}_j\boldsymbol{d}_j^\top / \lVert \boldsymbol{d}_j \rVert_2^2$	Logits-as-Features Model

Table 1: Modeling choices for $\boldsymbol{H}$, where $\boldsymbol{d}_j = \boldsymbol{e}_j - \frac{1}{K}\boldsymbol{1}_p$ for $j \in [K]$.

We estimate the LE matrix under ***different modeling choices* of the similarity matrix H, as shown in Table 1**, from simulations on GeoMNIST, a dataset consisting of three geometric shapes, using a variant of the AlexNet. We show in Figures 1 the estimated tail index r_α ($-\ln\alpha(t) \approx C\text{-}r_\alpha \ln t$) versus the corrupted label ratio p_{err}. As p_{err} increases, we expect LE effects diminish since the dataset is becoming more like random data.

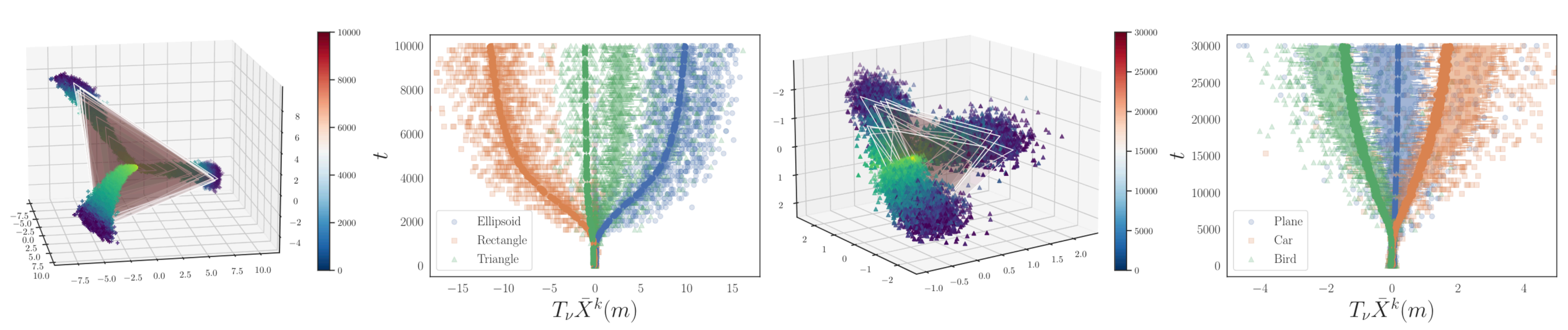
- In the I-model (Figure 1(a)), a clear phase transition for the tail of (α - β) occurs around $p_{err} = 2/3$, when the dataset has completely random labels.
- Although in the L-model (Figure 1(b)), the phase transition for the tail is less obvious, note around $p_{err} = 2/3$ the index of β begins to dominate that of α.

Corollary: Connection with Neural Collapse

Neural collapse[4-5] is a recent phenomenological finding on the geometry of logits of DNNs at convergence: they tend to form *equiangular tight frames* (ETFs). Let $B(t)$ be the definite integral of $\beta(\tau)$ from $\tau = 0$ to $\tau = t$, we have the following corollary.

Proposition (Neural Collapse of the LE-ODE)

Under L-model and the same setup as in Theorem 1, if $\gamma(t) > 0$ and there exists some $T > 0$ such that $B(t) < 0$ for $t \geq T$, then $\left\{\bar{\boldsymbol{X}}^k(t)/\lVert\bar{\boldsymbol{X}}^k(t)\rVert\right\}_{k=1}^K$ forms an ETF as $t \to \infty$.



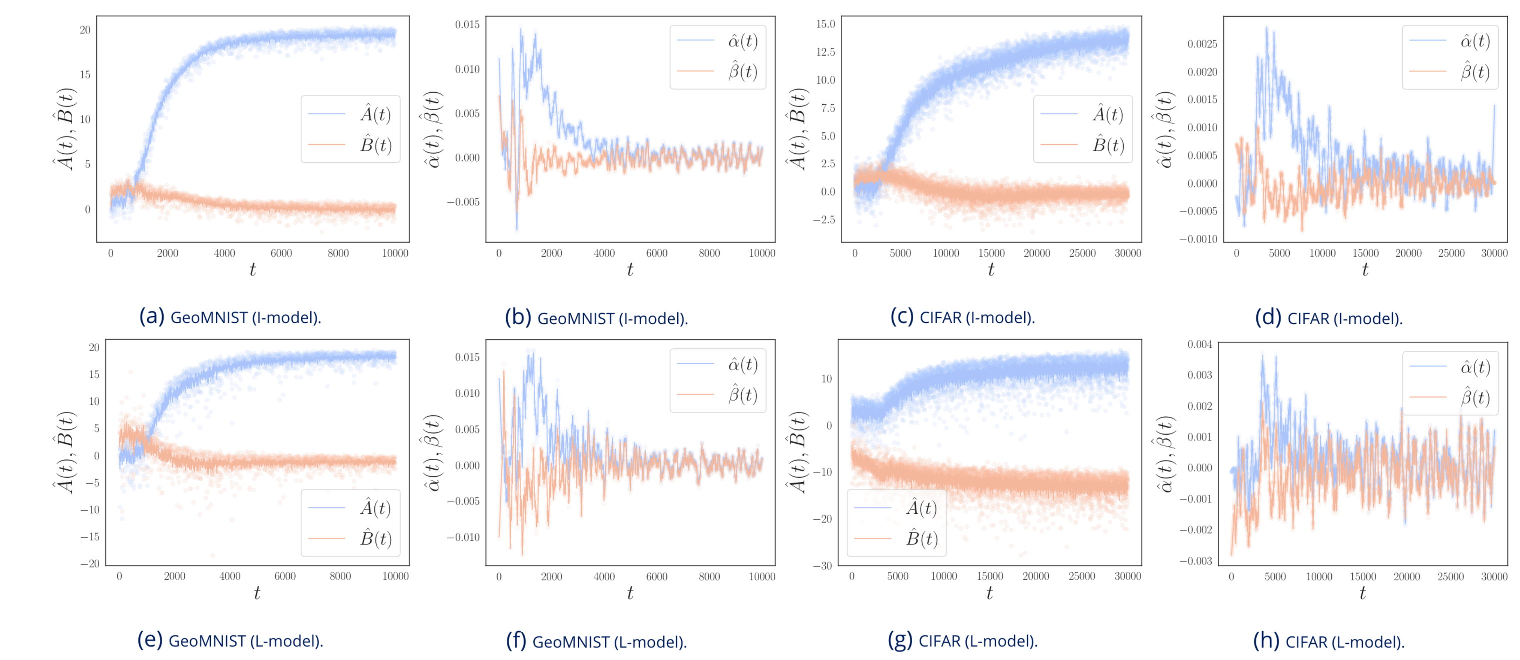
(a) GeoMNIST in $\mathbb{R}^3$. (b) GeoMNIST in $\mathbb{R}$. (c) CIFAR in $\mathbb{R}^3$. (d) CIFAR in $\mathbb{R}$.

Figure 2: Logit evolution in training. More details are given in the caption of Figure 1 of our paper. Here note that the logits are tending to form an ETF as training progresses (c.f. (a) and (c)).

Imitating DNN Dynamics

Estimating the LE Matrix

Under the I-model and the L-model, the solutions to the LE-ODE can be solved exactly, whence we are able to estimate the LE matrix in terms of the parameters $\alpha(t)$ and $\beta(t)$ based on their cumulative functions $A(t)$ and $B(t)$.



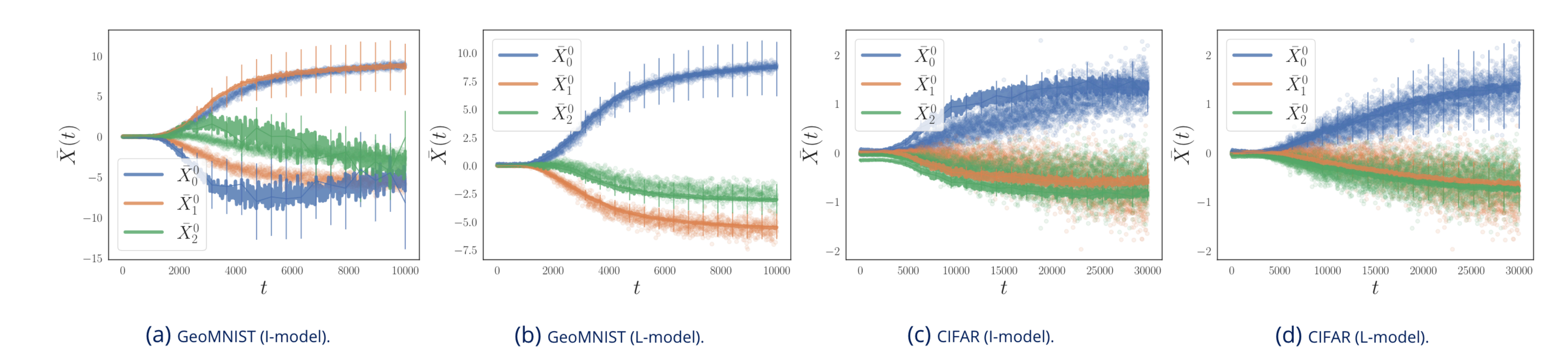
(a) GeoMNIST (I-model). (b) GeoMNIST (I-model). (c) CIFAR (I-model). (d) CIFAR (I-model).
(e) GeoMNIST (L-model). (f) GeoMNIST (L-model). (g) CIFAR (L-model). (h) CIFAR (L-model).

Figure 3: Estimated parameters for the LE matrix $\boldsymbol{E}(t)$.

Imitating DNN Dynamics and Evaluations

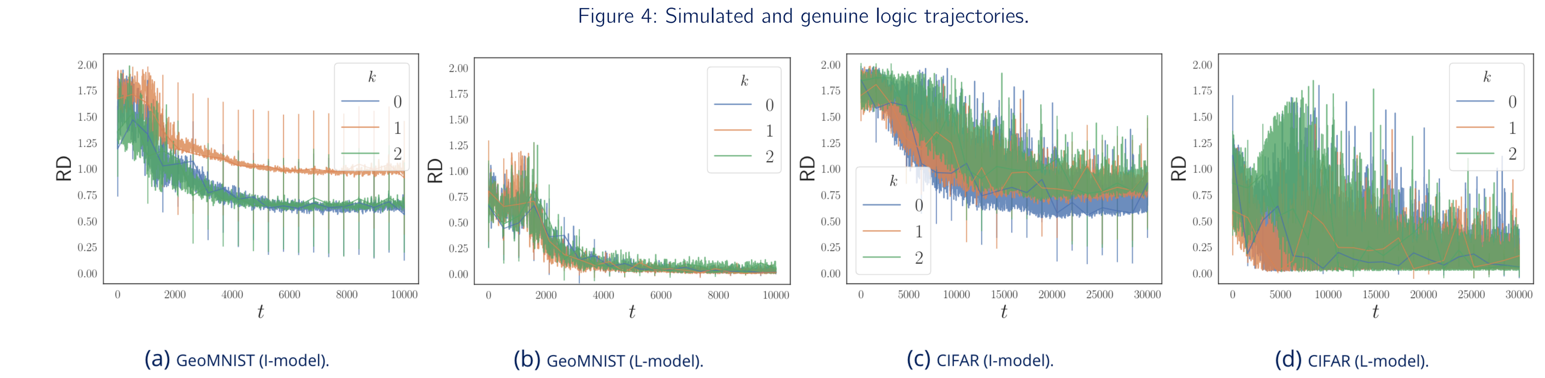
With $\boldsymbol{E}(t)$ estimated, we can simulate the LE-SDE using say forward Euler method to test if our model specification is reasonably correct. We assess the goodness-of-fit via the following relative difference (RD, the lower the better):

$$\mathrm{RD}_k(t) := \frac{\lVert \bar{\boldsymbol{X}}^k(t) - \bar{\boldsymbol{Y}}^k(t) \rVert_{\boldsymbol{H}^k}}{\left(\lVert \bar{\boldsymbol{X}}^k(t) \rVert_2 + \lVert \bar{\boldsymbol{Y}}^k(t) \rVert_2\right)/2}.$$



(a) GeoMNIST (I-model). (b) GeoMNIST (L-model). (c) CIFAR (I-model). (d) CIFAR (L-model).

Figure 4: Simulated and genuine logic trajectories.

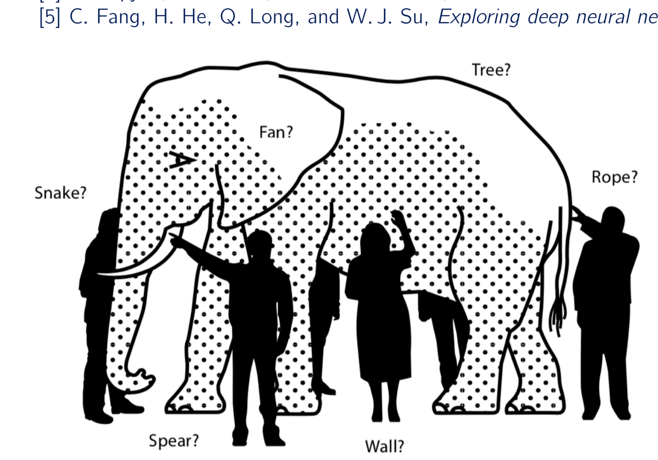


(a) GeoMNIST (I-model). (b) GeoMNIST (L-model). (c) CIFAR (I-model). (d) CIFAR (L-model).

Figure 5: Relative difference (RD) between simulated and genuine trajectories.

References

[1] H. He and W. J. Su, *The local elasticity of neural networks*, ICLR, 2020.
[2] S. Chen, H. He, W. J. Su, *Label-aware neural tangent kernel: Toward better generalization and local elasticity*, NeurIPS, 2020.
[3] Z. Deng, H. He, W. J. Su, *Toward better generalization bounds with locally elastic stability*, ICML, 2021.
[4] V. Papyan, X. Y. Han, and D. L. Donoho, *Prevalence of neural collapse during the terminal phase of deep learning training*, PNAS, 2020.
[5] C. Fang, H. He, Q. Long, and W. J. Su, *Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training*, PNAS, 2021.



[Video] [Slides] [Paper]