The Prevalence of Neural Collapse in Neural Multivariate Regression

G. Andriopoulos, Z. Dong, L. Guo, Z. Zhao, K. Ross

12.12.2024

NEURAL INFORMATION PROCESSING SYSTEMS

NEW YORK UNIVERSITY

جامعة نيويورك أبوظبي
NYU ABU DHABI

NYU 上海 SHANGHAI 纽约大学

PART 01

Motivation

Motivation

Neural Collapse (NC): observed during TPT of large overparameterized models for classification.

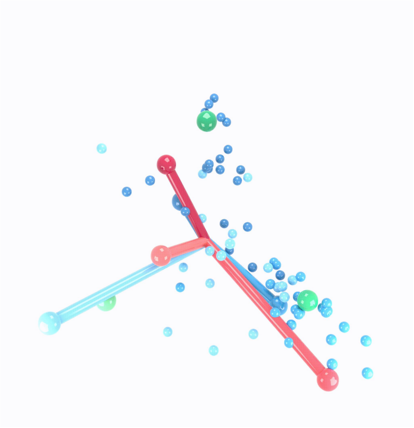
TPT: The post zero-error training phase

Papyan et al. (2020) outlined properties that describe the emergence of a geometric structure that induces maximally separated clustering between last-layer features and linear classifiers:

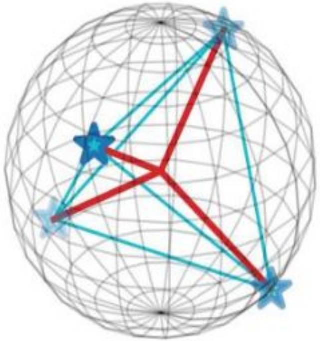
- **NC1 Variability Collapse**
- **NC2 Convergence to Simplex ETF**
- **NC3 Convergence to Self-Duality**
- **NC4 Nearest Class-Mean Decision Rule**

Empirical observations of NC were coupled by theoretical frameworks such as the unconstrained feature model (UFM).

The **UFM** helps explain why NC occurs in classification by allowing the optimization to freely adjust last-layer features along with classifier weights.



Papyan et al., 2020



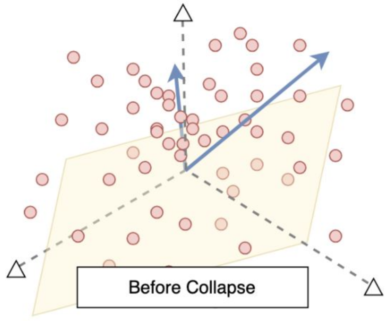
Kim et al., 2024

The figure

Motivation

Recently,

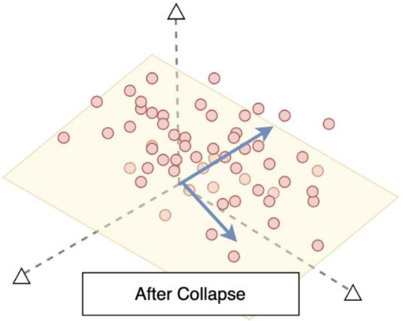
- NC has been investigated under different ***loss functions and regularization techniques***.
- NC properties have been examined within ***intermediate layers of DNNs***.
- NC phenomena have been studied for both ***balanced/imbalanced data scenarios***.
- Under the NC framework, criteria have been devised for the ***detection of OOD data***.
- NC provided a theoretical framework, which explained the ***bias-variance alignment*** in modern deep models.



The prevalence and implications of NC in regression remained unexplored.

Regression serves numerous applications across diverse domains such as:

- Imitation learning for autonomous driving.
- Robotics.
- Forecasting stock prices, estimating risk, and predicting market trends.
- Meteorology.
- RL algorithms, where regression is employed to predict value functions, with the targets being Monte Carlo or bootstrapped returns.



Our work introduces **Neural Regression Collapse (NRC)** as a new form of NC for **neural multivariate regression**.

PART 02

Neural Regression Collapse (NRC)

Notations and Definitions

Notations

- **Multivariate regression**: $\{(\mathbf{x}_i, \mathbf{y}_i), i = 1, ..., M\}$
- **Targets**: **n-dim** with sample **cov matrix** Σ and **min eigenvalue** $\lambda_{\min}$
- **DNN**: $f_{\theta,\mathbf{W},\mathbf{b}}(\mathbf{x}) = \mathbf{W}\mathbf{h}_\theta(\mathbf{x}) + \mathbf{b}$
- Non-linear **feature extractor**: $\mathbf{h}_\theta() : \mathbb{R}^D \to \mathbb{R}^d, \mathbf{h}_i := \mathbf{h}_\theta(\mathbf{x}_i), \widetilde{\mathbf{h}}_i := \mathbf{h}_i \cdot ||\mathbf{h}_i||^{-1}$
- **Feature matrix**: $\mathbf{H} := [\mathbf{h}_1 \cdots \mathbf{h}_M]$
- **Final linear layer**: $\mathbf{W}$
- For most neural regression tasks: **n << d**
- Train the DNN using GD to minimize the regularized L2 loss:

$$\min_{\theta,\mathbf{W},\mathbf{b}} \frac{1}{2M} \sum_{i=1}^{M} ||f_{\theta,\mathbf{W},\mathbf{b}}(\mathbf{x}_i) - \mathbf{y}_i||_2^2 + \frac{\lambda_\theta}{2}||\theta||_2^2 + \frac{\lambda_{\mathbf{W}}}{2}||\mathbf{W}||_F^2$$

Definitions

Additional notation:

- $\mathrm{proj}(\mathbf{v}|\mathbf{C})$: **projection** of v to the **column space** of C.
- $\mathbf{H}_{PCA_n}$: columns consisting of the **first n principal components** of the features.

NRC1:

- ***Feature vector collapse***
- The d-dim feature vectors collapse to a n-dim subspace spanned by their n principal components:

$$\text{NRC1} = \frac{1}{M}\sum_{i=1}^{M}\left|\left|\tilde{\mathbf{h}}_i - proj(\tilde{\mathbf{h}}_i|\mathbf{H}_{PCA_n})\right|\right|_2^2 \to 0.$$

NRC2:

- ***Self duality***
- The feature vectors also collapse to the n-dim space spanned by the rows of the last-layer weight matrix:

$$\text{NRC2} = \frac{1}{M}\sum_{i=1}^{M}\left|\left|\tilde{\mathbf{h}}_i - proj(\tilde{\mathbf{h}}_i|\mathbf{W}^T)\right|\right|_2^2 \to 0.$$

NRC3:

- The Gram matrix of the last-layer weights converges to a specific functional form that depends on the square root of the covariance matrix of the targets. There exists a constant $\gamma \in (0, \lambda_{\min})$, such that:

$$\text{NRC3} = \left|\left|\frac{\mathbf{W}\mathbf{W}^T}{||\mathbf{W}\mathbf{W}^T||_F} - \frac{\mathbf{\Sigma}^{1/2} - \gamma^{1/2}\mathbf{I}_n}{||\mathbf{\Sigma}^{1/2} - \gamma^{1/2}\mathbf{I}_n||_F}\right|\right|_F^2 \to 0.$$

PART 03

Main Experiments

Prevalence of NRC in Practice

Experiment Setup

- **MuJoCo Locomotion Datasets** (Reacher, Swimmer, Hopper) [**Brockman et al., 2016**]:
  - Simulated robotic locomotion tasks with continuous control environments.
  - **Input**: State observations of the robot
  - **Target**: Optimal actions to achieve a task
  - **Model**: Multi-Layer Perceptron with 3 hidden layers of dimension 256.

- **CARLA Dataset** [**Dosovitskiy et al., 2017**]:
  - Autonomous driving simulation with diverse traffic and environmental conditions.
  - **Input**: RGB images from vehicle-mounted cameras.
  - **Target**: Steering commands for navigation, e.g. speed and angle
  - **Model**: ResNet18

- **UTK Face Dataset** [**Zhang et al., 2017**]:
  - Large-scale facial dataset labeled with age, gender, and ethnicity.
  - **Input**: Facial images
  - **Target**: Predicted attributes, e.g. age
  - **Model**: ResNet 34

Results: Prevalence of NRC1-3

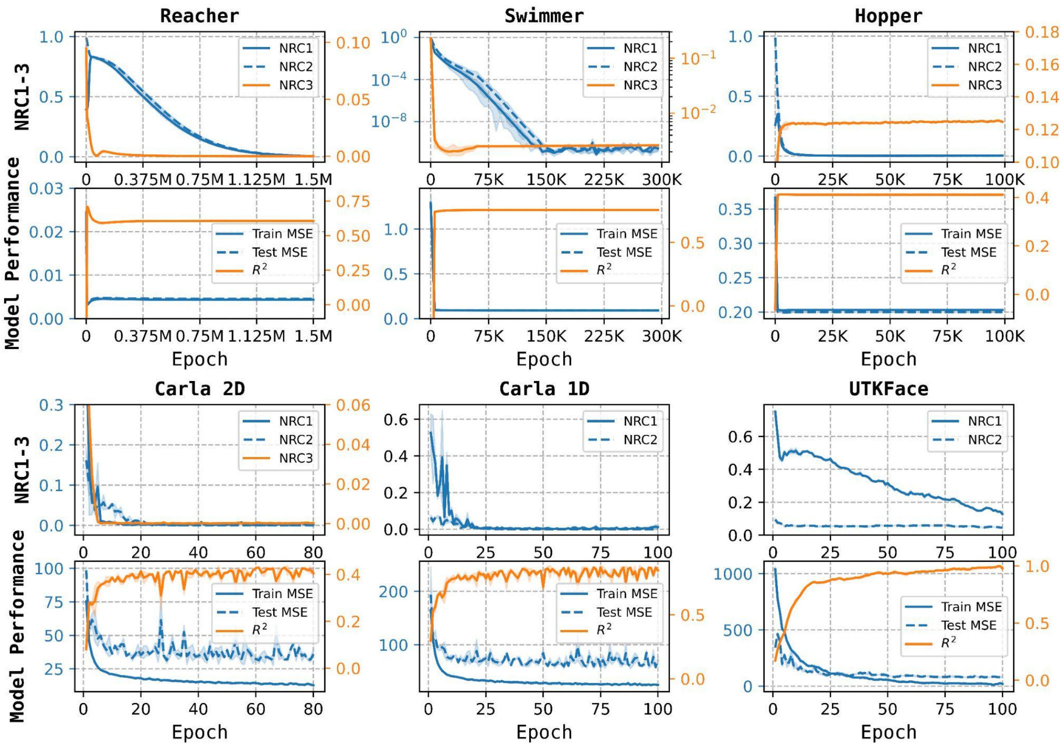


Figure 1: Prevalence of NRC1-NRC3 in the six datasets. Model performances are also shown.

- Converging model performance metrics indicate training becomes stable.
- The presence of NRC1-NRC3 across six datasets indicates that neural collapse is not only prevalent in classification but also often occurs in multivariate regression.

NEURAL INFORMATION PROCESSING SYSTEMS

NEW YORK UNIVERSITY

جامعة نيويورك أبوظبي
NYU ABU DHABI

NYU 上海 SHANGHAI 纽约大学

Results (Cont.): Explained Variance Ratio

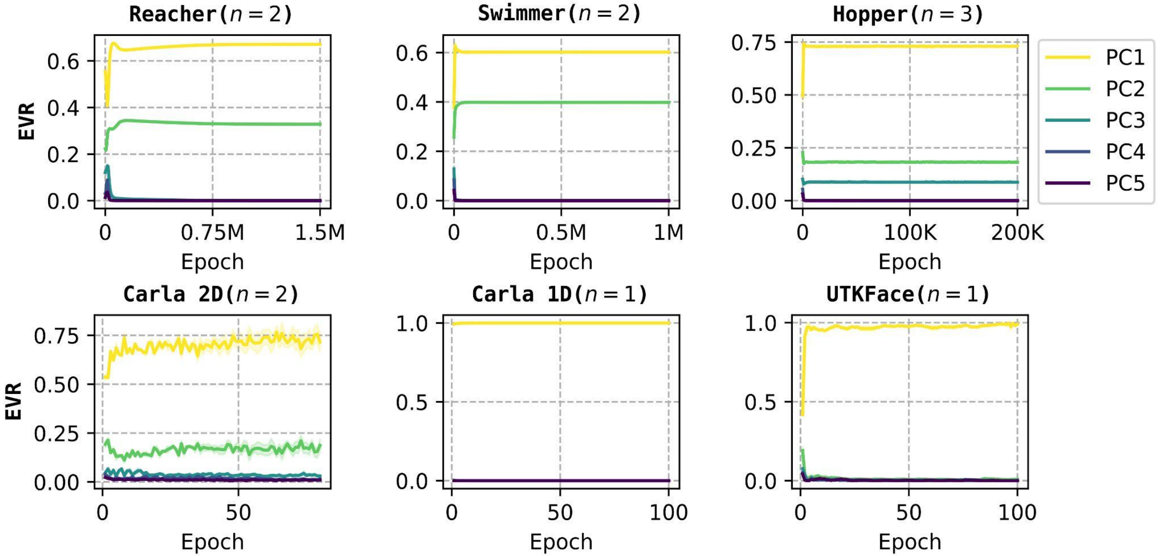


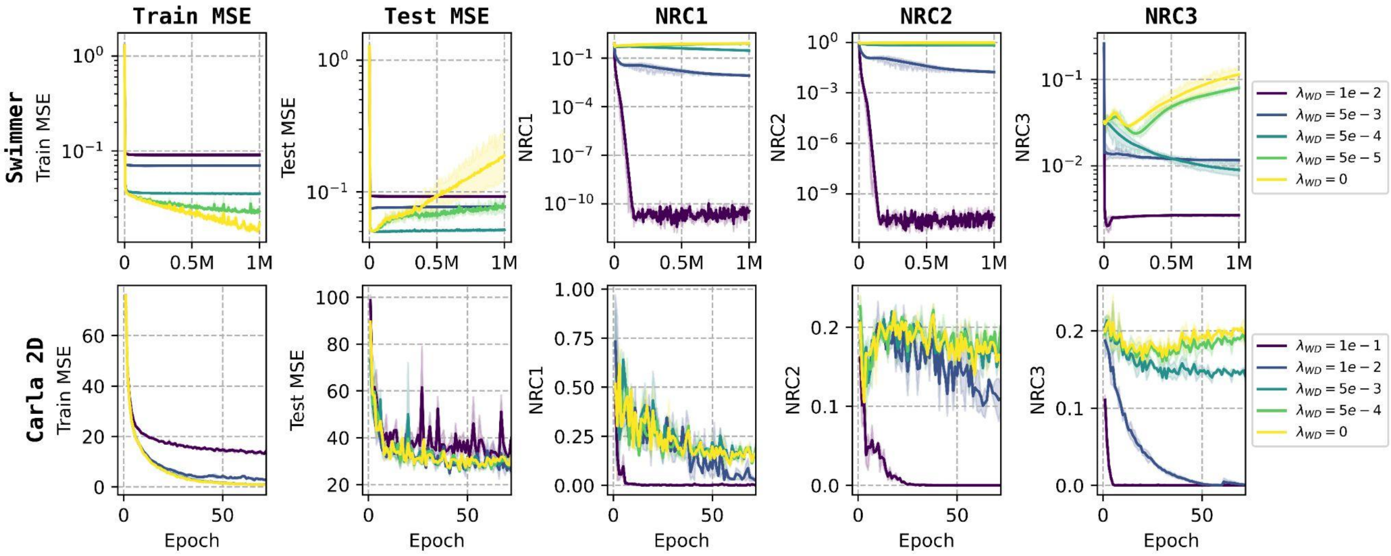
Figure 2: Explained Variance Ratio (EVR) for the first 5 principal components (PC) of **H** during training. Target dimension is denoted as *n*.

- Significant variance for all of the **first *n* components** after a short period of training;
- Very low or even no variance for other components;
- A perfect collapse occurs in the subspace spanned by the first *n* principal components.

NEURAL INFORMATION PROCESSING SYSTEMS

NEW YORK UNIVERSITY
جامعة نيويورك أبوظبي
NYU ABU DHABI
NYU 上海 SHANGHAI 纽约大学

Results (Cont.): Small Weight Decays



- As weight decay decreases, NRC1-3 become less;
- NRC1-3 that emerges during training is **due to regularization**

Figure 3: Examine NRC1-NRC3 with different weight decay values.

PART 04

Theoretic Results

Prevalence of NRC in Theory

Theoretic results

NRC1-3 emerge as solutions in the **regularized UFM**:

$$\frac{1}{2M}||\mathbf{WH} + \mathbf{b}\mathbf{1}_M^T - \mathbf{Y}||_F^2 + \frac{\lambda_{\mathbf{H}}}{2M}||\mathbf{H}||_F^2 + \frac{\lambda_{\mathbf{W}}}{2}||\mathbf{W}||_F^2$$

- All of the d-dim feature vectors lie in the n-dim space spanned by the n rows of $\mathbf{W}$:

$$\mathbf{H} = \sqrt{\frac{\lambda_{\mathbf{W}}}{\lambda_{\mathbf{H}}}}\mathbf{W}^T[\mathbf{\Sigma}^{1/2}]^{-1}(\mathbf{Y} - \bar{\mathbf{Y}})$$

- The theoretical result matches the definition of **NRC3** for $c = \lambda_{\mathbf{W}}\lambda_{\mathbf{H}}$:

$$\mathbf{W}\mathbf{W}^T = \sqrt{\frac{\lambda_{\mathbf{H}}}{\lambda_{\mathbf{W}}}}\left[\mathbf{\Sigma}^{1/2} - \sqrt{c}\mathbf{I}_n\right]$$

- At optimality, the **residual errors** are **uncorrelated** across the n target dimensions and each has **variance equal to c**:

$$\mathbf{WH} + \mathbf{b}\mathbf{1}_M^T - \mathbf{Y} = -\sqrt{c}[\mathbf{\Sigma}^{1/2}]^{-1}(\mathbf{Y} - \bar{\mathbf{Y}})$$

No regularization implies **no collapse**: the emergence of NRC is due to inclusion of regularization in the loss function.

References

1. Papyan, Han, Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences, 117(40): 24652-24663, 2020.*
2. Fang, He, Long, Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalance training. *Proceedings of the National Academy of Sciences, 118(43): e2103091118, 2021.*
3. Mixon, Parshall, Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis, 20(2): 11, 2022.*
4. Zhou, Li, Ding, You, Qu, Zhu. On the optimization landscape of neural collapse under MSE loss: Global optimality with unconstrained features. *International Conference on Machine Learning, pages 27179-27202. PMLR, 2022a.*
5. Guo, Ross, Zhao, Andriopoulos, Ling, Xu, Dong. Cross entropy versus label smoothing: A neural collapse perspective. *arxiv preprint arxiv: 2402.03979, 2024.*
6. Galanti, György, Hutter. On the role of neural collapse in transfer learning. *arXiv preprint arXiv:2112.15121, 2021.*
7. Brockman, Cheung, Pettersson, Schneider, Schulman, Tang, and Zaremba. Openai gym, 2016.
8. Dosovitskiy, Ros, Codevilla, López and Koltun. CARLA: An Open Urban Driving Simulator. *Conference on Robot Learning (2017).*
9. Zhang, Song and Qi. Age Progression/Regression by Conditional Adversarial Autoencoder. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017): 4352-4360.