NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2025

The Thirty-Ninth Annual Conference on Neural Information Processing Systems

Neural Collapse in Cumulative Link Models for Ordinal Regression: An Analysis with Unconstrained Feature Model

Chuang Ma[1], Tomoyuki Obuchi[1,2], Toshiyuki Tanaka[1]

[1]Graduate School of Informatics, Kyoto University
[2]RIKEN Center for Advanced Intelligence Project (AIP)

@ San Diego Convention Center, San Diego, California, USA
Dec 2 – 7, 2025

RIKEN AIP Informatics KYOTO UNIVERSITY FOUNDED 1897 KYOTO JAPAN

京都大学
KYOTO UNIVERSITY

Background: Neural Collapse & Unconstrained Feature Model

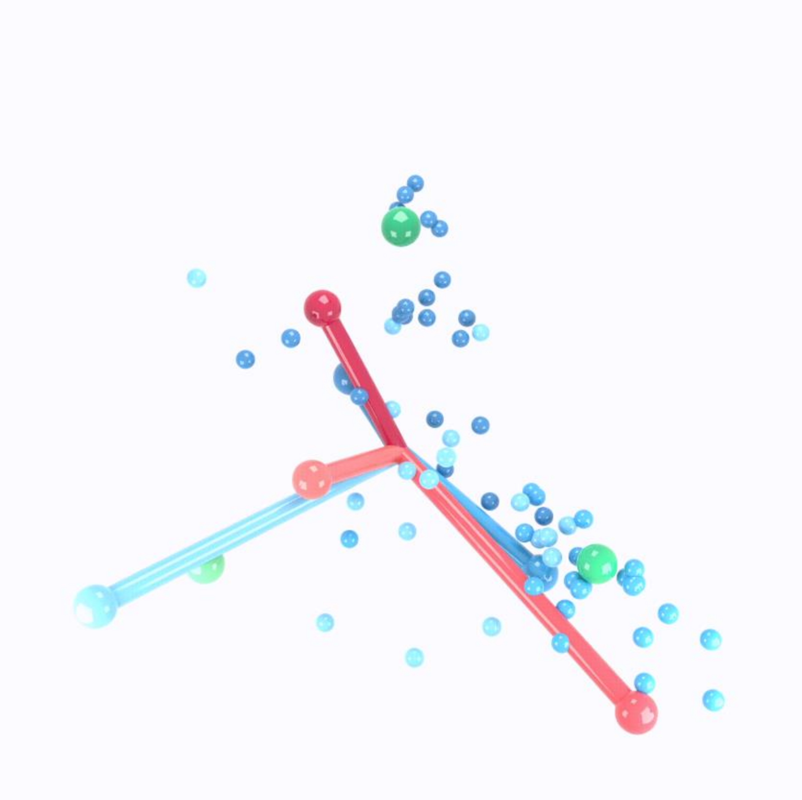
Finding of Neural Collapse (In classification tasks on balanced datasets)

After sufficient training, features of the penultimate layer and the final classifier weights in sufficiently expressive DNNs exhibit a remarkably simple symmetric structure [Papyan, Han, Donoho (2020)].

- **(NC1)**: Within-class mean collapse
- **(NC2)**: Convergence to simplex Equiangular Tight Frame (ETF)
- **(NC3)**: Convergence to self-duality
- **(NC4)**: The network simply classifies by nearest class mean

Visualization of Neural Collapse →

- **Green spheres**: the vertices of the standard Simplex ETF
- **Red ball-and-sticks**: linear classifiers
- **Blue ball-and-sticks**: feature class means
- **Small blue spheres**: last-layer features



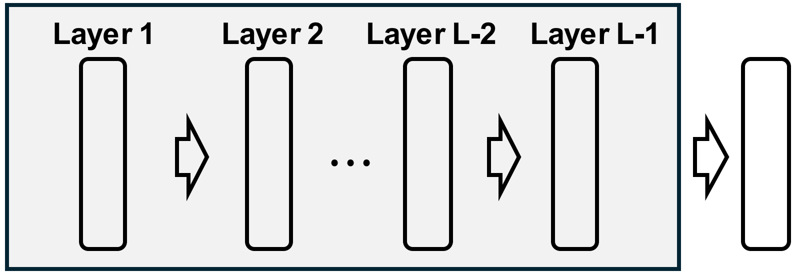
Animation from [Papyan, Han, Donoho (2020)]

Unconstrained Feature Model (UFM)

Theoretical Model for NC

- Unconstrained Feature Model (UFM) [Mixon et al. 2020]
- Layer-Peeled Model (LPM) [Fang et al. 2021]

1-layer UFM:



- **Real Neural Network**

$$(\boldsymbol{\theta}^*, \boldsymbol{W}^*) = argmin_{\boldsymbol{\theta},\boldsymbol{W}} \frac{1}{N} \sum_{c=1}^{C} \sum_{i}^{N_c} \mathcal{L}\left(\boldsymbol{W}, \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}_{c,i})\right) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2 + \frac{\lambda}{2} \|\boldsymbol{W}\|^2$$

- **UFM/LPM of 1-layer**

$$(\boldsymbol{H}^*, \boldsymbol{W}^*) = argmin_{\boldsymbol{H},\boldsymbol{W}} \frac{1}{N} \sum_{c=1}^{C} \sum_{i}^{N_c} \mathcal{L}(\boldsymbol{W}, \boldsymbol{h}_{c,i}) + \frac{\lambda_H}{2N} \|\boldsymbol{H}\|^2 + \frac{\lambda_W}{2} \|\boldsymbol{W}\|^2$$

$$\boldsymbol{H} = \left(\boldsymbol{h}_{c,i}\right)_{c=1,\ldots,C,i=1,\ldots,N_c}$$

- $\mathcal{L}$ can be any loss. E.g., Cross entropy (CE), Mean square error (MSE).

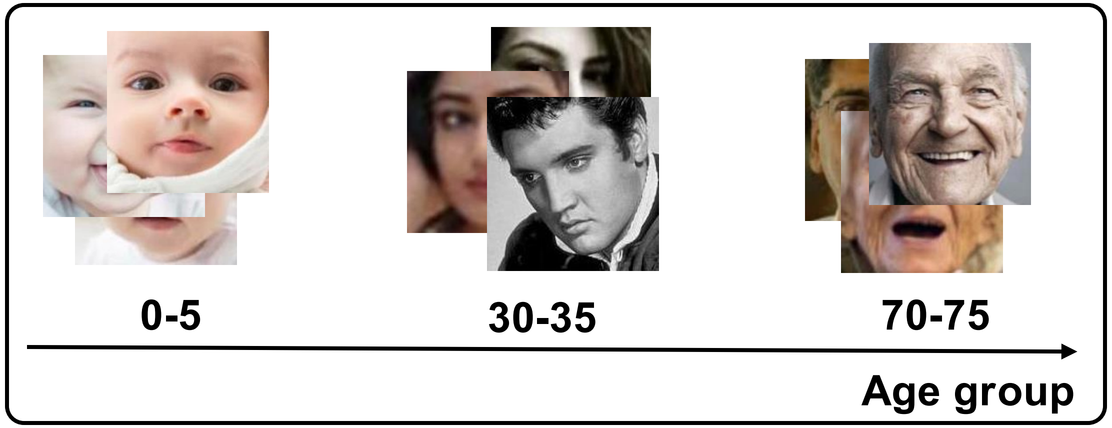
Related work: NC extensions beyond classification

[Andriopoulos et al., 2024] generalized NC to multivariate regression to find NRC. In addition, the concept of NC has also been extended to many settings such as multi-label classification [Li et al., 2024], LLMs [Wu and Papyan, 2024], diffusion models [Nguyen et al., 2024] and transfer learning [Galanti et al., 2022, Li et al., 2024].

Ordinal regression (OR)

- Discrete labels with a natural order (unlike standard classification).
- Greater distance ⇒ larger loss.
    - For examples, **Age estimation**, mistaking 25 as 40 is worse than as 30, etc.

An example of age estimation dataset



Our Contribution

Extend NC to OR

Ordinal regression (OR)

Formulation

- **Dataset:**
  - Input space $\mathcal{X}$, ordered label set $\mathcal{Y} = \{1,2,\ldots,Q\}$ $(1 < 2 < \cdots < Q)$
  - $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}, (\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$
  - $\mathcal{D}_q = \{(\boldsymbol{x}_i, y_i) \in \mathcal{D} | y_i = q\}_{i=1}^{N}$, $\mathrm{n_q} = |\mathrm{D_q}|$, $\sum_{\mathrm{q=1}}^{\mathrm{Q}} \mathrm{n_q} = \mathrm{N}$
- **Object:** learn a function $\mathcal{X} \longrightarrow \mathcal{Y}$ from $\mathcal{D}$
- **Cumulative Link Model (CLM):**
  - Introduce a latent variable $z \in \mathbb{R}$, associate class labels with partitions of the z -axis.
  - Thresholds $\boldsymbol{b} = (b_0 \ b_1, \ldots, b_Q)$ with $b_0 < b_1 < \cdots < b_Q$: $y = q \Leftrightarrow z \in (b_{q-1}, b_q]$
  - Using an inverse link function $g: \mathbb{R} \rightarrow (0,1)$ to relate y and z :

$$P(y \leq q \mid z) = g(b_q - z) \Leftrightarrow$$
$$P(y = q \mid z) = g(b_q - z) - g(b_{q-1} - z)$$

  - Standard choice: logistic function $g(x) = (1 + e^{-x})^{-1}$, etc.
  - Learn a mapping $z = f(x_i)$ from input x_i to the latent scalar z.
    - Using DNN as f: $z = f_{\boldsymbol{w},\theta}(\boldsymbol{x}_i) = \boldsymbol{w}^\top \boldsymbol{h}_\theta(\boldsymbol{x}_i)$ [Vargas et al., 2020]

Theoretical analysis based on UFM

- **Maximum Likelihood Estimation with regularization** $(N = |\mathcal{D}|)$

$$\min_{\boldsymbol{w},\theta}\left\{-\frac{1}{N}\sum_{q}\sum_{i\in\mathcal{D}_q}\log\Big(g\big(b_q-\boldsymbol{w}^\top\boldsymbol{h}_\theta(\boldsymbol{x}_i)\big)-g\big(b_{q-1}-\boldsymbol{w}^\top\boldsymbol{h}_\theta(\boldsymbol{x}_i)\big)\Big)+R(\theta)+\frac{1}{2}\lambda_w\|\boldsymbol{w}\|_2^2\right\}$$

- **Corresponding UFM** $(n_q = |\mathcal{D}_q|, \sum_q n_q = N)$

$$\min_{\boldsymbol{w},\boldsymbol{H}}\left\{-\frac{1}{N}\sum_{q}\sum_{i=1}^{n_q}\left\{\log\Big(g\big(b_q-\boldsymbol{w}^\top\boldsymbol{h}_{q,i}\big)-g\big(b_{y_i-1}-\boldsymbol{w}^\top\boldsymbol{h}_{q,i}\big)\Big)+\frac{1}{2}\lambda_h\left\|\boldsymbol{h}_{q,i}\right\|_2^2\right\}+\frac{1}{2}\lambda_w\|\boldsymbol{w}\|_2^2\right\}$$

→ Analyzing this UFM enables us to study NC in CLM-based OR

Ordinal Neural Collapse (ONC)

ONC is characterized by the following three properties:

- ONC1 (Within-class mean collapse)

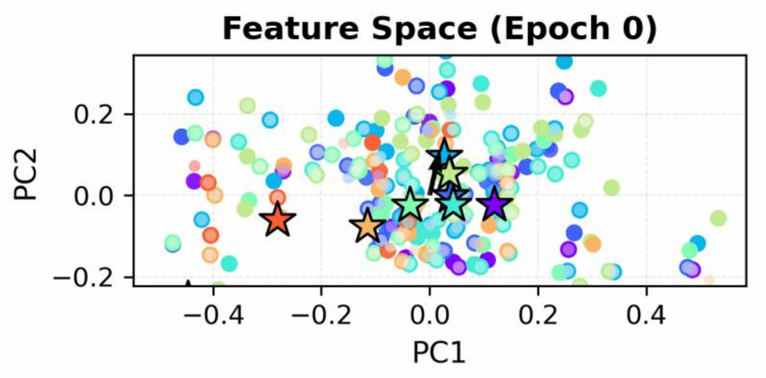
$$\boldsymbol{h}_{q,i}^* = \boldsymbol{h}_q^*$$

- ONC2 (Self-duality)

$$\boldsymbol{h}_q^* \parallel \boldsymbol{w}^*, \forall q$$

- ONC3 (latent variable alignment)

$$z_q^* = (\boldsymbol{w}^*)^\top \boldsymbol{h}_q^* \Rightarrow z_1 \leq z_2 \leq \cdots z_Q$$



Animated visualization of feature space evolution during training on a real dataset.

How to determine $w^* = \|\boldsymbol{w}^*\|_2, z_q^*$?

Optimality condition of UFM

$$\frac{g'(b_q - z_q^*) - g'(b_{q-1} - z_q^*)}{g(b_q - z_q^*) - g(b_{q-1} - z_q^*)} + \lambda_h \frac{z_q^*}{(w^*)^2} = 0$$

$$\lambda_w w^* - \frac{\lambda_h}{(w^*)^3} \sum_{q=1}^{Q} \alpha_q (z_q^*)^2 = 0 \quad \left(\alpha_q = \frac{n_q}{N}\right)$$

We prove the result under the assumption that g is differentiable and g' is log concave.

Some Limits and Phase Transition (1/2)

Particularly interesting phenomena

- Phase transition from nontrivial ($w^*, z_q^* \neq 0$) to trivial ($w^* = z_q^* = 0$) solution
    - Phase boundary: $\lambda_w \lambda_h = C \equiv \sum_{q=1}^{Q} \alpha_q \left(\frac{g\prime(b_q) - g\prime(b_{q-1})}{g(b_q) - g(b_{q-1})} \right)^2$
- Simple and local behavior in the vanishing regularization limit ($\lambda_w \times \lambda_h \to 0$)
  - z_q^* is determined by $g'(b_q - z_q^*) = g'(b_{q-1} - z_q^*)$
    - In the symmetric g ($1 - g(x) = g(-x)$) such as logistic function, this means

$$z_q^* = \frac{b_q + b_{q-1}}{2}.$$

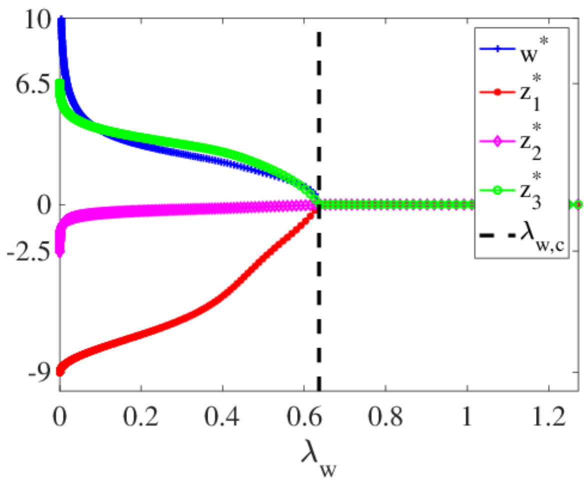
  - In the limit,

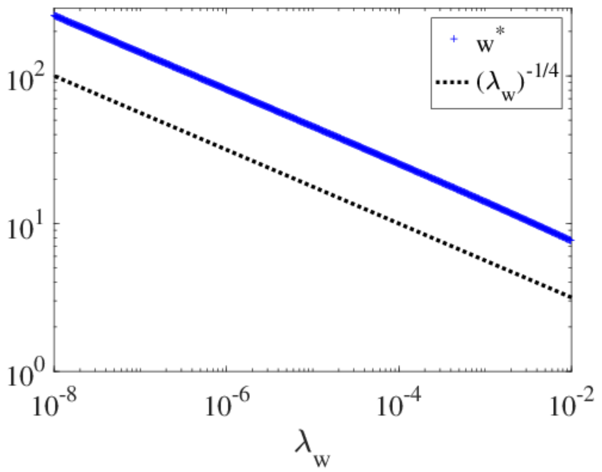
$$w^* = O\left(\left(\frac{\lambda_h}{\lambda_w}\right)^{1/4}\right).$$

  Thus it may diverge, vanish, and remain finite depending on how the limit is taken.

Some Limits and Phase Transition (2/2)

Numerical Example





Solution behavior of EOS in the logit model for $Q = 3$ with $b = (-10, -8, 3, 10)$ at $\lambda_h = 1$.

(Left) w^* and z^* are plotted against λ_w on a linear scale.

- A clear phase transition appears a $\lambda_{w,c} = C/\lambda_h$(vertical broken line)
- The values of z^* in the limit $\lambda_w \to 0$ match well with the theoretical prediction $\left(z_q^* = \left(b_q + b_{q-1}\right)/2\right)$.

(Right) w^* is plotted on a log-log scale in the small-λ_w region.

- A power-law divergence with exponent $-1/4$, corresponding to the scaling $w^* = O\left((\lambda_h/\lambda_w)^{1/4}\right)$ with fixed λ_h, is clearly observed.

Experiments (1/3) - Setup

Examine ONC actually happens in DNN experiments

- **Dataset:**
  - Tabular datasets: Publicly available dataset from [Gutiérrez et al., 2016]. Five largest datasets in the site (ER, LE, SW, CA, and WR) are used.
  - Image dataset: UTKFace age estimation dataset [Zhang et al., 2017] grouped ages into classes with five-year intervals..
- **DNNs:**
  - For tabular datasets: we employed a multilayer perceptron with residual connections.
  - For image dataset: we used ResNet101 and ResNet50 [He et al., 2016], and DenseNet201 [Huang et al., 2017] as backbones.

- **DNN setup:** a very weak regularization.

- **Treatment of thresholds:**
  - Fixed thresholds : b_Q is set to a large positive value. $b_0 = -b_Q$.
    Others are evenly spaced over $[b_0, b_Q]$
  - Learnable case: $b_Q = \infty, b_0 = -\infty$, other parameters are optimized by the ML method.
    No regularization applied.

Experiments (2/3) - Evaluation metrics

$\bar{\boldsymbol{h}}_q = \frac{1}{n_q}\sum_{i\in\mathcal{D}_q} \boldsymbol{h}_\theta(\boldsymbol{x}_i), \bar{\boldsymbol{h}} = \frac{1}{N}\sum_{i\in D} \boldsymbol{h}_\theta(\boldsymbol{x}_i)$, $\boldsymbol{u}$: the 1st principal component of $\{\bar{\boldsymbol{h}}_q - \bar{\boldsymbol{h}}_q\}$

$$\underline{\mathrm{ONC}_1} = \frac{(1/Q)\sum_{q=1}^{Q} \frac{1}{N_q} \sum_{(\boldsymbol{x}_i, y_i)\in D_q} \|\boldsymbol{h}_\theta(\boldsymbol{x}_i) - \bar{\boldsymbol{h}}_q\|_2}{(1/N)\sum_{i=1}^{N} \|\boldsymbol{h}_\theta(\boldsymbol{x}_i) - \bar{\boldsymbol{h}}\|_2},$$

$$\underline{\mathrm{ONC}_{2\text{-}1}} = \frac{\sum_{q=1}^{Q} \|(\bar{\boldsymbol{h}}_q - \bar{\boldsymbol{h}}) - (\boldsymbol{u}^\top(\bar{\boldsymbol{h}}_q - \bar{\boldsymbol{h}}))\,\boldsymbol{u}\|_2^2}{\sum_{q=1}^{Q} \|\bar{\boldsymbol{h}}_q - \bar{\boldsymbol{h}}\|_2^2}, \quad \underline{\mathrm{ONC}_{2\text{-}2}} = 1 - \left|\frac{\boldsymbol{w}^\top \boldsymbol{u}}{\|\boldsymbol{w}\|_2}\right|,$$

$$\underline{\mathrm{ONC}_3} = \frac{\sum_{q=1}^{Q-1} |b_q - (z_q + z_{q+1})/2|}{\sum_{q=1}^{Q-1} (b_{q+1} - b_q)},$$

Recall

$$\frac{g'(b_q - z_q^*) - g'(b_{q-1} - z_q^*)}{g(b_q - z_q^*) - g(b_{q-1} - z_q^*)} + \lambda_h \frac{z_q^*}{(w^*)^2} = 0$$

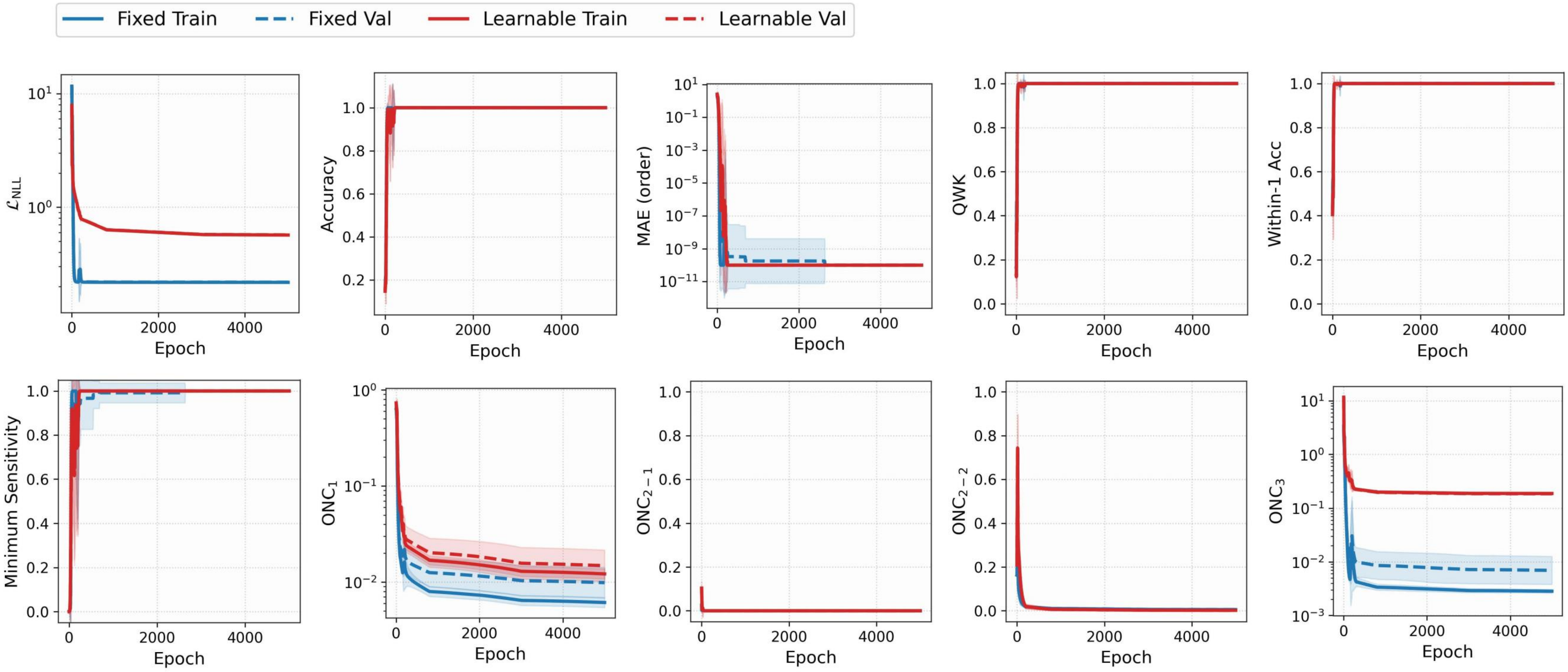
$$\lambda_w w^* - \frac{\lambda_h}{(w^*)^3} \sum_{q=1}^{Q} \alpha_q (z_q^*)^2 = 0 \quad \left(\alpha_q = \frac{n_q}{N}\right)$$

In the vanishing regularization limit ($\lambda_w \times \lambda_h \to 0$)

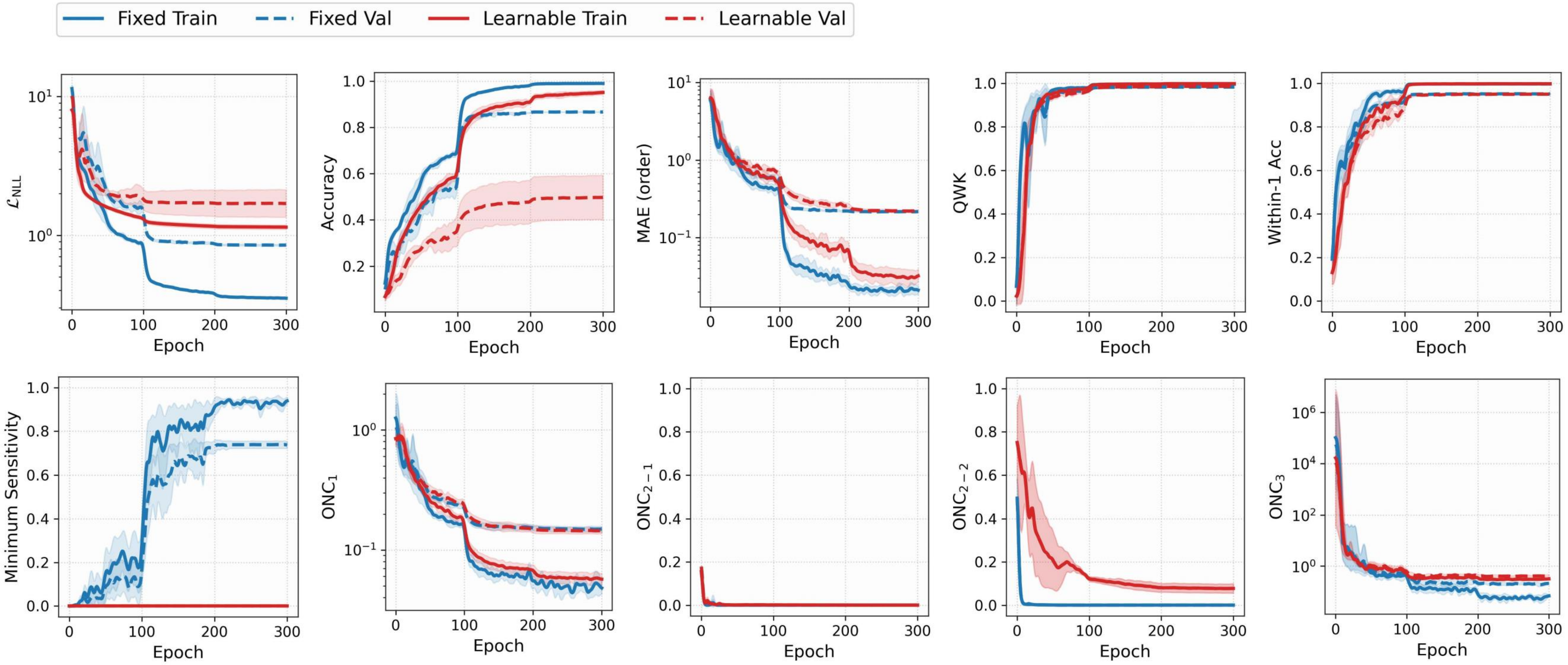
- In the **symmetric link function:**

$$\boldsymbol{z_q^* = \frac{b_q + b_{q-1}}{2}}$$

Experiments (2/3) - Evaluation metrics

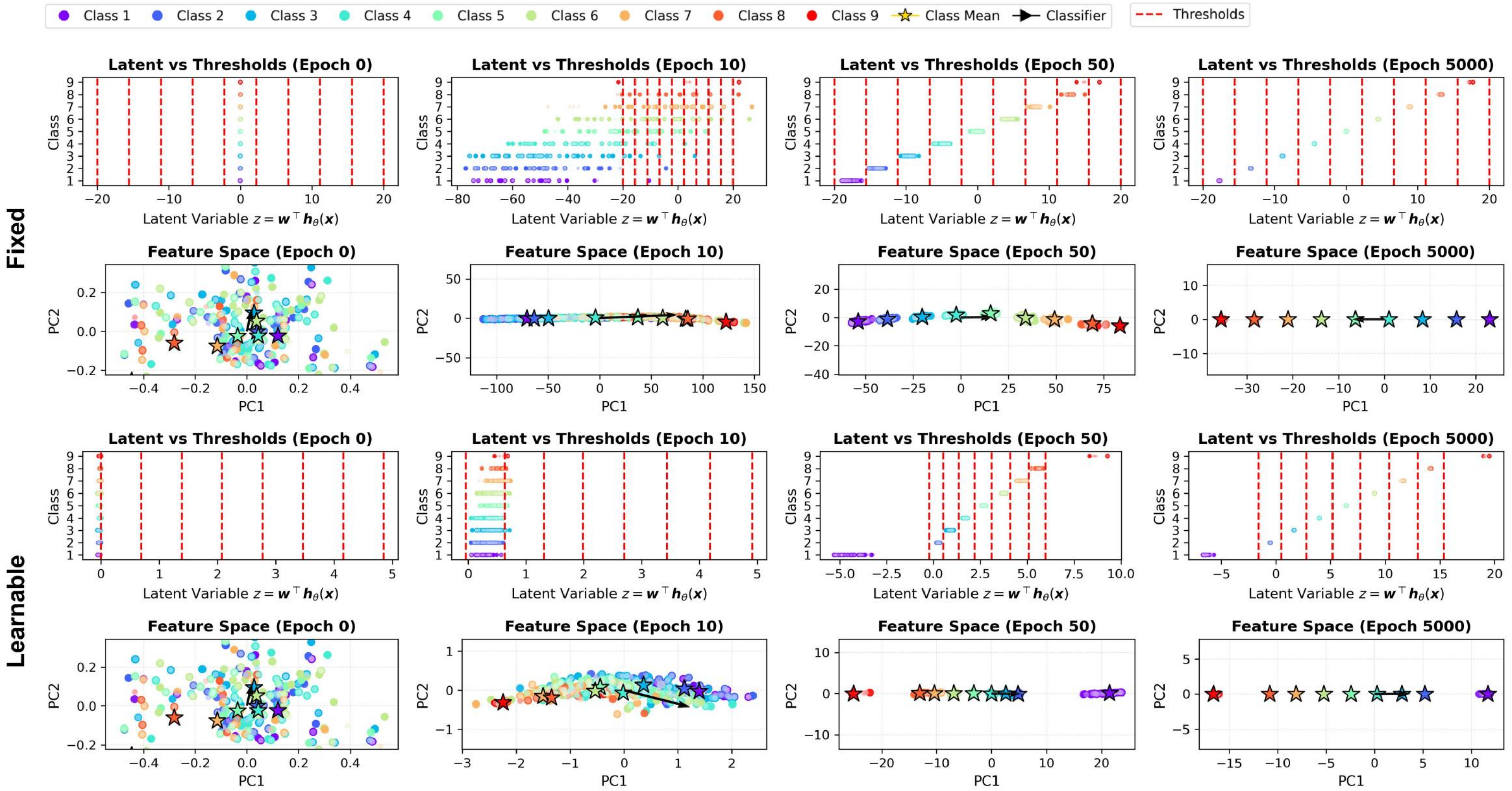


Epoch-wise average metrics curves for the ER dataset with the logit model.



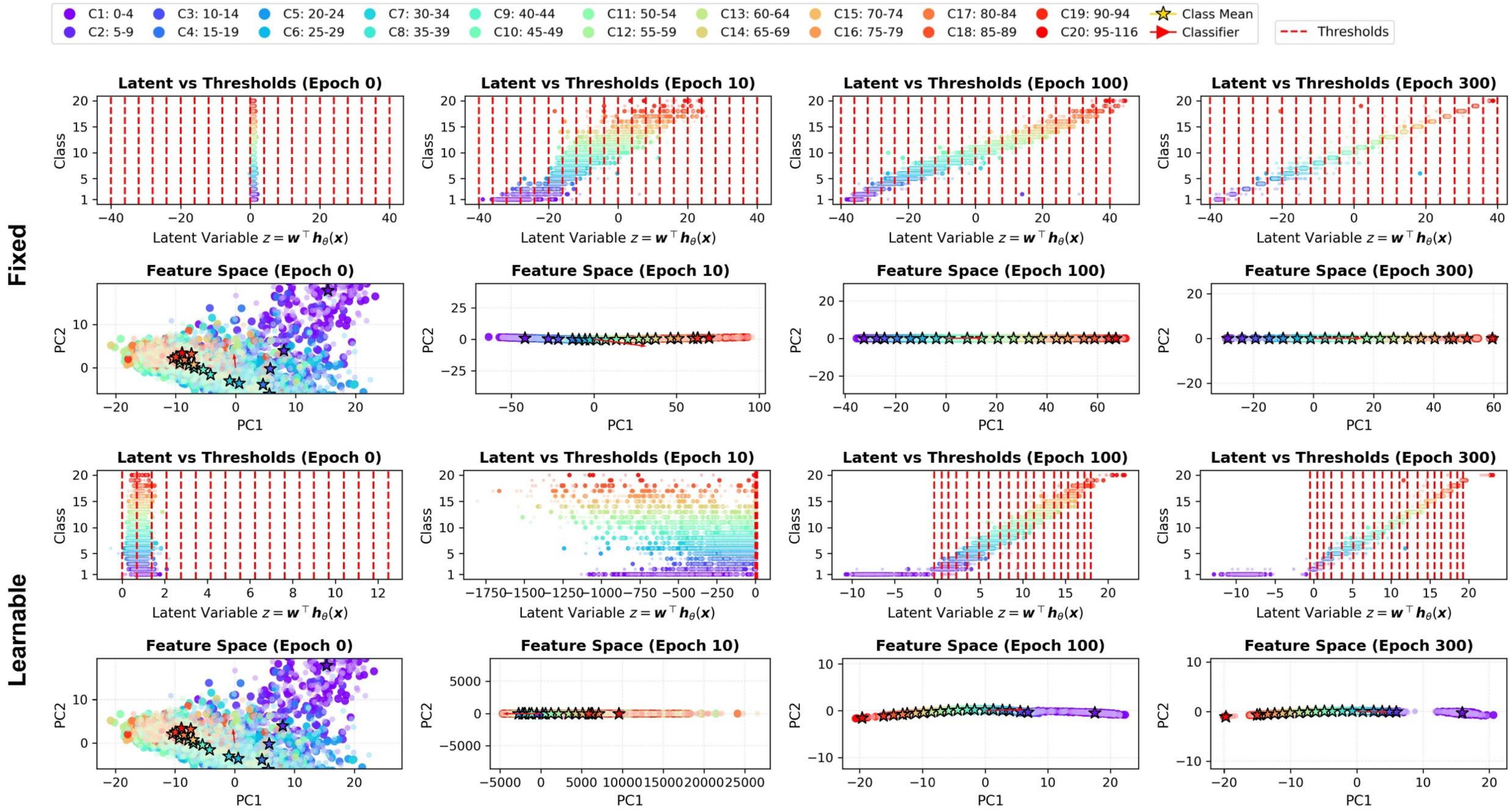
Epoch-wise average metrics curves for the UTKFace dataset with ResNet101 backbone and logit model.

Experiments (3/3) - Visualization



Latent and feature space visualization for the ER dataset with the logit model.

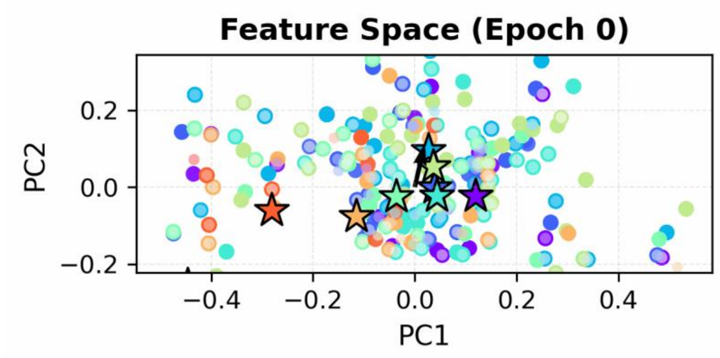
Experiments (3/3) - Visualization



Latent and feature space visualization for the UTKFace dataset with ResNet101 backbone and logit model.

Summary and Perspectives

- **Summary: Extended NC to CLM-based OR**
  - A UFM invention for OR is the key.
  - Found ONC
    - $\mathrm{ONC1}$: within-class mean collapse
    - $\mathrm{ONC2}$: self-duality
    - $\mathrm{ONC3}$: latent variable alignment
      - Particularly simple in the weak regularization limit.
  - ONC happens in real DNN experiments.



- **Perspectives**
  - New loss/regularization may be proposed from ONC
    - Imbalance problems will be mitigated.
    - Convergence is expected to be accelerated.

References (1/2)

Papyan, Vardan, X. Y. Han, and David L. Donoho. "Prevalence of neural collapse during the terminal phase of deep learning training." Proceedings of the National Academy of Sciences 117.40 (2020): 24652-24663.

Fang C, He H, Long Q, et al. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training[J]. Proceedings of the National Academy of Sciences, 2021, 118(43): e2103091118.

Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. Sampling Theory, Signal Processing, and Data Analysis, 20(2):11, 2022.

George Andriopoulos, Zixuan Dong, Li Guo, Zifan Zhao, and Keith Ross. The prevalence of neural collapse in neural multivariate regression. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 126417-126451, 2024.

Pengyu Li, Xiao Li, Yutong Wang, and Qing Qu. Neural collapse in multi-label learning with pick-all-label loss. In ICML'24: Proceedings of the 41th International Conference on Machine Learning, volume 235, pages 28060-28094. PMLR, 2024a.

Robert Wu and Vardan Papyan. Linguistic collapse: Neural collapse in (large) language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 137432-137473, 2024.

Bac Nguyen, Chieh-Hsin Lai, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Stefano Ermon, and Yuki Mitsufuji. Mitigating embedding collapse in diffusion models for categorical data. arXiv preprint arXiv:2410.14758v1 [cs.LG], 2024.

References (2/2)

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In International Conference on Learning Representations 2022, 2022.

Xiao Li, Sheng Liu, Jinxin Zhou, Xinyu Lu, Carlos Fernandez-Granda, Zhihui Zhu, and Qing Qu. Understanding and improving transfer learning of deep models via neural collapse. Transactions on Machine Learning Research, May 2024b.

Victor Manuel Vargas, Pedro Antonio Gutiérrez, and César Hervás-Martínez. Cumulative link models for deep ordinal classification. Neurocomputing, 401:48-58, August 2020.

P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez. Ordinal regression methods: Survey and experimental study. IEEE Transactions on Knowledge and Data Engineering, 28(1):127-146, January 2016. doi: 10.1109/TKDE.2015.2457911.

Zhifei Zhang, Yang Song, and Hairong Qi. Age Progression/Regression by Conditional Adversarial Autoencoder . In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4352-4360, Los Alamitos, CA, USA, July 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.463.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770-778, 2016. doi: 10.1109/CVPR.2016.90.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2261-2269, 2017. doi: 10.1109/CVPR.2017.243.