Perturbation Analysis of Neural Collapse

Bar-Ilan University

Tom Tirer*, Haoxiang Huang* & Jonathan Niles-Weed

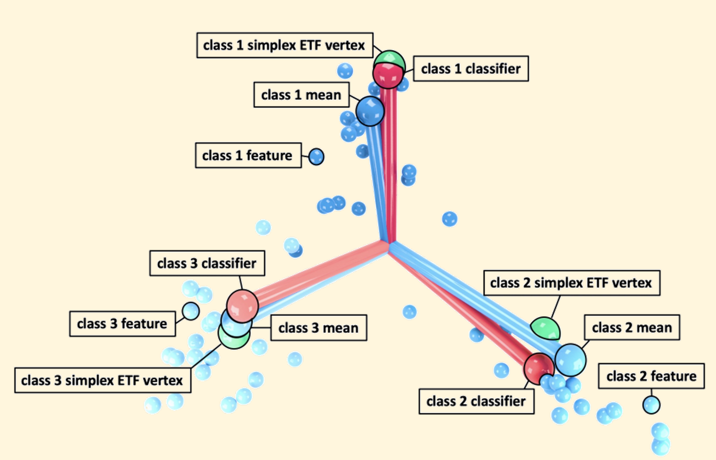
New York University

The Neural Collapse (NC) Phenomenon

- DNN-based classifiers (of K classes) can be typically represented as
$$\boldsymbol{\psi}_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}) + \boldsymbol{b}$$
where $\boldsymbol{x} \in \mathbb{R}^D$ is the sample, $\boldsymbol{h}_{\boldsymbol{\theta}}(\cdot): \mathbb{R}^D \to \mathbb{R}^d$ is the (deep) feature mapping, and $\{\boldsymbol{W} \in \mathbb{R}^{K\times d}, \boldsymbol{b} \in \mathbb{R}^K\}$ is the last layer classifier. Learnable params: $\boldsymbol{\Theta} = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{\theta}\}$.
- Common practice: Keep optimizing the network's parameters after the training error vanishes to further push the training loss toward zero.
- The "Neural Collapse" (NC) phenomenon [Papyan et al. (2020)] has been empirically observed in this phase of training with CE loss (or MSE loss [Han et al. (2022)]):
Let $\boldsymbol{H} \coloneqq [\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}_{1,1}), \dots, \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}_{1,n}), \dots \dots, \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}_{K,1}), \dots, \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}_{K,n})] \in \mathbb{R}^{d\times Kn}$.
  - (NC1): Decrease in within-class variability of features $\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x})$:
  $\|\boldsymbol{H} - \overline{\boldsymbol{H}} \otimes \mathbf{1}_n^\top\|_F$ decreases, where $\overline{\boldsymbol{H}} \coloneqq [\overline{\boldsymbol{h}}_1, \dots, \overline{\boldsymbol{h}}_K] \in \mathbb{R}^{d\times K}$ are classes' mean features
  - (NC2): Increase in the similarity of the mean features to a simplex ETF structure:
  $\left\|(\overline{\boldsymbol{H}} - \overline{\boldsymbol{h}}_G \mathbf{1}_K^\top)^\top(\overline{\boldsymbol{H}} - \overline{\boldsymbol{h}}_G \mathbf{1}_K^\top) - \rho\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\right\|_F$ decreases, for some $\rho > 0$
  - (NC3): Increase in the alignment of the last weights $\boldsymbol{W}^\top$ and the mean features $\overline{\boldsymbol{H}}$:
  $\left\|\boldsymbol{W}(\overline{\boldsymbol{H}} - \overline{\boldsymbol{h}}_G \mathbf{1}_K^\top) - \tilde{\rho}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\right\|_F$ decreases, for some $\tilde{\rho} > 0$

Empirical observations in practical settings:

- "NC metrics" typically plateau above zero (even when reducing LR)
- The margin from exact NC depends on the dataset complexity, DNN architecture, hyperparameter tuning, etc.
- Interesting depthwise behavior: gradual reduction of within-class variability (NC1 metric)



The Unconstrained Features Model (UFM)

- The typical way to optimize the DNN's parameters (empirical risk minimization):
$$\min_{\boldsymbol{\Theta}} \frac{1}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n} \mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}_{k,i}) + \boldsymbol{b}, \boldsymbol{y}_k) + \mathcal{R}(\boldsymbol{\Theta})$$
where $\boldsymbol{y}_k \in \mathbb{R}^K$ is one-hot vector, $\mathcal{L}(\cdot,\cdot)$ is a loss function (e.g., CE or MSE) and $\mathcal{R}(\cdot)$ is a regularization term (e.g., squared ℓ_2-norm)
- [Mixon et al. (2020)] suggested to explore NC via the Unconstrained Features Model (UFM) – the features $\{\boldsymbol{h}_{k,i} \coloneqq \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}_{k,i})\}$ are free optimization variables:
$$\min_{\boldsymbol{W},\boldsymbol{b},\{\boldsymbol{h}_{k,i}\}} \frac{1}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n} \mathcal{L}(\boldsymbol{W}\boldsymbol{h}_{k,i} + \boldsymbol{b}, \boldsymbol{y}_k) + \mathcal{R}(\boldsymbol{W}, \boldsymbol{b}, \{\boldsymbol{h}_{k,i}\})$$
- Most (if not all) of the existing theoretical works on NC consider UFM settings.
The typical result: All the minimizers exhibit exact NC structures (zero NC metrics) with no effect of regularization hyperparameters on the structure
- UFMs limitations: cannot explain the aforementioned observations

This Work Is About:

- Exploiting knowledge on gradient dynamics and minimizers of UFMs for studying practical (non-exact) NC behavior.

Existing and New UFM Results

Theorem 3.1 in [Tirer & Bruna, 2022] (characterization of minimizers)

Let $d \geq K$, $c \coloneqq \sqrt{\lambda_H \lambda_W}$ and $\rho \coloneqq \max\{(1-c)\sqrt{\lambda_W/\lambda_H}, 0\}$. Any global minimizer $(\boldsymbol{W}^*, \boldsymbol{H}^*)$ of
$$\min_{\boldsymbol{W}\in\mathbb{R}^{K\times d}, \boldsymbol{H}\in\mathbb{R}^{d\times Kn}} \mathcal{L}(\boldsymbol{W}, \boldsymbol{H}) \coloneqq \frac{1}{2Kn}\|\boldsymbol{W}\boldsymbol{H} - \boldsymbol{Y}\|_F^2 + \frac{\lambda_W}{2K}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2Kn}\|\boldsymbol{H}\|_F^2$$
obeys that $\boldsymbol{H}^* = \overline{\boldsymbol{H}} \otimes 1_n^\top$ for some $\overline{\boldsymbol{H}} \coloneqq [\boldsymbol{h}_1^*, \dots, \boldsymbol{h}_K^*] \in \mathbb{R}^{d\times K}$, $\boldsymbol{W}^{*\top} = \sqrt{\lambda_H/\lambda_W}\boldsymbol{H}$, and
$$\overline{\boldsymbol{H}}^\top\overline{\boldsymbol{H}} = \rho \boldsymbol{I}_K \implies (\overline{\boldsymbol{H}} - \boldsymbol{h}_G^* 1_K^\top)^\top(\overline{\boldsymbol{H}} - \boldsymbol{h}_G^* 1_K^\top) = \rho(\boldsymbol{I}_K - \frac{1}{K}1_K 1_K^\top)$$

- **New & useful NC1 metric:** $\boxed{\widetilde{NC}_1(\boldsymbol{H}) \coloneqq \mathrm{trace}(\boldsymbol{\Sigma}_W(\boldsymbol{H}))/\mathrm{trace}(\boldsymbol{\Sigma}_B(\boldsymbol{H}))}$
$\boldsymbol{\Sigma}_W(\boldsymbol{H})$ and $\boldsymbol{\Sigma}_B(\boldsymbol{H})$ are the within- and between-class covariance matrices
- More amenable for theoretical analysis than $NC_1(H) \coloneqq \frac{1}{K}\mathrm{trace}(\boldsymbol{\Sigma}_W(\boldsymbol{H})\boldsymbol{\Sigma}_B^\dagger(\boldsymbol{H}))$
- For fixed $\boldsymbol{H}$, the minimizer w.r.t. $\boldsymbol{W}$: $\boldsymbol{W}^*(\boldsymbol{H}) = \boldsymbol{Y}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{H}^\top + n\lambda_W \boldsymbol{I}_d)^{-1}$
- [Han et al. (2022)] empirically showed that $\|\boldsymbol{W}\boldsymbol{H} - \boldsymbol{Y}\|_F^2 - \|\boldsymbol{W}^*(\boldsymbol{H})\boldsymbol{H} - \boldsymbol{Y}\|_F^2$ is small during MSE minimization of practical DNN classifiers

Theorem (NC1 metric decreases along the gradient flow)

Assume that $\lambda_W > 0$, $\lambda_H \geq 0$, and that $\boldsymbol{H}_0$ is non-collapsed (i.e., $\boldsymbol{\Sigma}_W(\boldsymbol{H}_0) \neq 0$). Then, along the gradient flow: $\frac{d\boldsymbol{H}_t}{dt} = -Kn\nabla\mathcal{L}(\boldsymbol{W}^*(\boldsymbol{H}_t), \boldsymbol{H}_t)$

- $\widetilde{NC}_1(\boldsymbol{H}_t)$ strictly decreases along the flow until it reaches zero.
- $t \mapsto e^{2\lambda_H t}\mathrm{trace}(\boldsymbol{\Sigma}_W(\boldsymbol{H}_t))$ decreases along the flow.
In particular, when $\lambda_H > 0$, $\mathrm{trace}(\boldsymbol{\Sigma}_W(\boldsymbol{H}_t))$ decays exponentially.
- $t \mapsto e^{2\lambda_H t}\mathrm{trace}(\boldsymbol{\Sigma}_B(\boldsymbol{H}_t))$ strictly increases along the flow.

- We got with minimal assumptions : separation between the behavior of $\boldsymbol{\Sigma}_W$ and $\boldsymbol{\Sigma}_B$ along the flow, $\widetilde{NC}_1 \to 0$ exponentially if $\lambda_H > 0$,

New Model: Constraining the UFM

$$\min_{\boldsymbol{W},\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H}; \boldsymbol{H}_0) \coloneqq \frac{1}{2Kn}\|\boldsymbol{W}\boldsymbol{H} - \boldsymbol{Y}\|_F^2 + \frac{\lambda_W}{2K}\|\boldsymbol{W}\|_F^2 + \frac{\lambda_H}{2Kn}\|\boldsymbol{H}\|_F^2 + \frac{\beta}{2Kn}\|\boldsymbol{H} - \boldsymbol{H}_0\|_F^2$$

- The $\beta \gg 1$ case: can be interpreted as simple architecture between $\boldsymbol{H}_0$ and $\boldsymbol{H}$ that significantly constrains $\boldsymbol{H}$ (e.g., $\boldsymbol{H}_0$ are features one layer before $\boldsymbol{H}$)
- Practical DL motivation for $\boldsymbol{H} \approx \boldsymbol{H}_0$: some ResNets, neural ODE, and DEQ

Corollary (Transferring orthogonal collapse properties from $\boldsymbol{H}_0$)

Let $d \geq K$, $\lambda_H\lambda_W < 1$, and let $(\boldsymbol{W}^*, \boldsymbol{H}^*)$ be a minimizer of $\mathcal{L}(\boldsymbol{W}, \boldsymbol{H})$. Then, the minimizer of $f(\boldsymbol{W}, \boldsymbol{H}; \boldsymbol{H}_0 = \boldsymbol{H}^*)$ is unique and it is given by $(\boldsymbol{W}^*, \boldsymbol{H}^*)$.

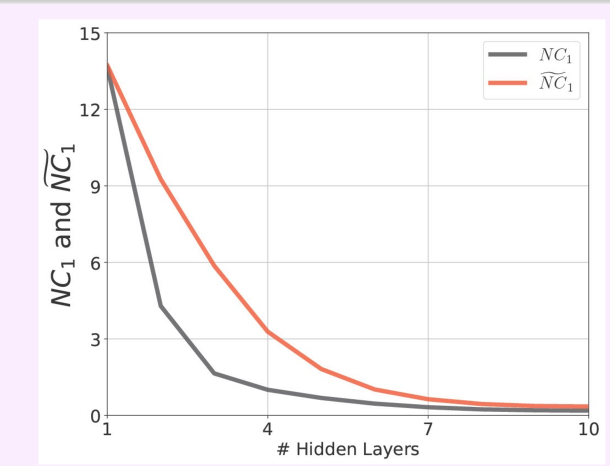
- Since we know a lot on $(\boldsymbol{W}^*, \boldsymbol{H}^*)$ minimizer of UFM — we can explore the near-collapse regime via perturbation analysis
- First order optimality condition:
$$\frac{\boldsymbol{H}_{1/\beta} - \boldsymbol{H}_0}{1/\beta} = -Kn\nabla\mathcal{L}(\boldsymbol{W}^*(\boldsymbol{H}_{1/\beta}), \boldsymbol{H}_{1/\beta})$$
where $\boldsymbol{H}_{1/\beta} = \min_{\boldsymbol{H}} f(\boldsymbol{W}^*(\boldsymbol{H}), \boldsymbol{H}; \boldsymbol{H}_0) = \min_{\boldsymbol{H}} \mathcal{L}(\boldsymbol{W}^*(\boldsymbol{H}), \boldsymbol{H}) + \frac{\beta}{2Kn}\|\boldsymbol{H} - \boldsymbol{H}_0\|_F^2$

Corollary (Depthwise decrease in NC1 – via gradient flow theory)

Assume that $\boldsymbol{H}_0$ is non-collapsed (i.e., $\boldsymbol{\Sigma}_W(\boldsymbol{H}_0) \neq 0$). For $\beta > C = C(\boldsymbol{H}_0)$, the minimizer of f, $\boldsymbol{H}_{1/\beta}$, obeys $\widetilde{NC}_1(\boldsymbol{H}_{1/\beta}) < \widetilde{NC}_1(\boldsymbol{H}_0)$.

- Numerical results:

Training an MLP on CIFAR-10 in layer-wise fashion (akin to updating $\boldsymbol{H}_0$ in our model with the previous $\boldsymbol{H}_{1/\beta}$)



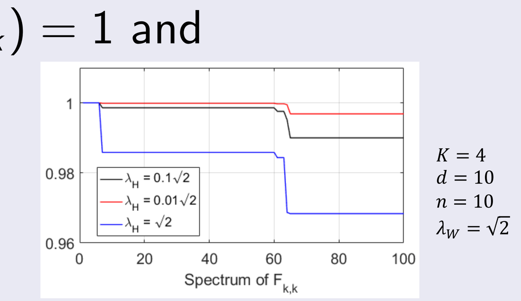
Analysis of the Near-Collapse Regime

Theorem (Perturbation analysis around collapse for $\beta \gg 1$)

Let $d > K$, $\lambda_H\lambda_W < 1$, and $\boldsymbol{H}_0 = \boldsymbol{H}^*$ where $(\boldsymbol{W}^*, \boldsymbol{H}^*)$ is a minimizer of $\mathcal{L}$ (i.e., collapsed). Set $\delta\boldsymbol{H}_0$, and let $(\tilde{\boldsymbol{W}}^*, \tilde{\boldsymbol{H}}^*)$ be the minimizer of $f(\cdot,\cdot;\tilde{\boldsymbol{H}}_0 = \boldsymbol{H}_0 + \delta\boldsymbol{H}_0)$.
Define $\delta\boldsymbol{H} \coloneqq \tilde{\boldsymbol{H}}^* - \boldsymbol{H}^*$.
For $\beta \gg \max\{1, \lambda_H\}$ we have (with approximation error of $O(\beta^{-2}, \|\delta\boldsymbol{H}_0\|^2)$)
$$\begin{bmatrix}\mathrm{vec}(\delta\boldsymbol{H}^{(1)}) \\ \vdots \\ \mathrm{vec}(\delta\boldsymbol{H}^{(K)})\end{bmatrix} \approx \begin{bmatrix}\boldsymbol{F}_{1,1} & \dots & \boldsymbol{F}_{1,K} \\ & \ddots & \\ \boldsymbol{F}_{K,1} & \dots & \boldsymbol{F}_{K,K}\end{bmatrix}\begin{bmatrix}\mathrm{vec}(\delta\boldsymbol{H}_0^{(1)}) \\ \vdots \\ \mathrm{vec}(\delta\boldsymbol{H}_0^{(K)})\end{bmatrix}$$
The $dn\times dn$ blocks have closed-form expressions made of $\boldsymbol{W}^*, \boldsymbol{H}^*, \lambda_W, \lambda_H, \beta$

Theorem (Spectral analysis of inter/intra class blocks)

Consider the setting of the previous theorem and let $k, \tilde{k} \in [K]$ with $k \neq \tilde{k}$.
We have that $\boldsymbol{F}_{k,k}$ is full rank, $\boldsymbol{F}_{k,\tilde{k}}$ is rank-1, $\sigma_{max}(\boldsymbol{F}_{k,k}) = 1$ and
$$\sigma_{min}(\boldsymbol{F}_{k,k}) = 1 - \beta^{-1}\sqrt{\lambda_H/\lambda_W}$$
$$\sigma_{max}(\boldsymbol{F}_{k,\tilde{k}}) = 2\beta^{-1}\lambda_H(1 - \sqrt{\lambda_H\lambda_W})$$
(*Actually, we compute the entire spectra)



- Insights gained from the model:
  - Increasing λ_H : increasing the intra-class (diagonal) blocks attenuation
  - Increasing λ_W : increasing the inter-class "interference" blocks attenuation
  - Main insight: the intra-class blocks (the effect of perturbation in a certain class in $\boldsymbol{H}_0$ on the features of the same class in $\boldsymbol{H}$) are the dominant. So λ_H plays the major role.
  - NC1 metric is less affected by the perturbations than other NC metrics (e.g., NC2)
- Numerical results: (*more results in the paper, including an "interference" study)

Training ResNet18 on CIFAR-10 with various weight decay (WD) settings – Modifying WD of feature mapping: more deviation from the baseline than modifying WD of last layer