Wide Neural Networks Trained with Weight Decay Provably Exhibit Neural Collapse

Minkyoung Song Jegwang Ryu

Jacot et al., ICLR 2025

2025.12.17

Overview

1. Preliminaries and Setup

2. Part 1: Conditions for Neural Collapse

3. Part 2: Gradient Descent Leads to NC

4. Numerical Results

5. Conclusion

Overview

1. **Preliminaries and Setup**
2. Part 1: Conditions for Neural Collapse
3. Part 2: Gradient Descent Leads to NC
4. Numerical Results
5. Conclusion

Key Contributions

Background. Deep neural networks at convergence are known to exhibit a symmetric geometric structure called Neural Collapse (NC).

Limitation of prior work. Most theoretical research has focused on the Unconstrained Features Model (UFM). While useful, the UFM treats features as free variables, which ignores the actual architecture of deep networks and the training data.

TL;DR

*This paper provides the first **end-to-end proof of Neural Collapse (NC)** in deep networks that end with **at least two linear layers**, specifically when trained via **Gradient Descent** with **Weight Decay**.*

Neural Collapse

Neural Collapse (NC)

- A pervasive phenomenon observed in the terminal phase of training [1].
- The geometric structure of the last-layer features collapses into a highly symmetric state.

Three Key Properties

- **NC1 (Within-Class Variability Collapse)**: The feature vectors of the same class collapse to a single mean vector.
- **NC2 (Simplex ETF)**: These class mean vectors form an Orthogonal or Simplex Equiangular Tight Frame.
- **NC3 (Alignment)**: The class means align with the weight vectors of the final layer.

[1] Papyan et al., Prevalence of neural collapse during the terminal phase of deep learning training, PNAS 2020.

Neural Collapse

Neural Collapse (NC)

- A pervasive phenomenon observed in the terminal phase of training [1].
- The geometric structure of the last-layer features collapses into a highly symmetric state.

Three Key Properties

- **NC1 (Within-Class Variability Collapse)**: The feature vectors of the same class collapse to a single mean vector.
- **NC2 (Simplex ETF)**: These class mean vectors form an Orthogonal or Simplex Equiangular Tight Frame.
- **NC3 (Alignment)**: The class means align with the weight vectors of the final layer.

[1] Papyan et al., Prevalence of neural collapse during the terminal phase of deep learning training, PNAS 2020.

Motivation

Problem. UFM is "Data-Agnostic" and mostly “Static” analysis.

- Data-Agnostic: Existing UFM theories treat features as free variables, ignoring the input data and backbone architecture.
- Static: They analyze *global minima*, ignoring how the network reaches them.
- **Question**: How do we prove NC in End-to-End deep networks?

Key Driver. Gradient Descent (GD) with Weight Decay

- We focus on GD with Weight Decay on Deep Networks.
  - **GD** drives *Interpolation* (Minimizing Training Error).
  - **Weight Decay** drives *Balancedness* between layers: $W_{l+1}^\top W_{l+1} \approx W_l W_l^\top$.

Problem Setup

Key Notations.

- Input/Output: Training data $X \in \mathbb{R}^{d \times N}$, One-hot labels $Y \in \mathbb{R}^{K \times N}$.
- Weights: $W_l \in \mathbb{R}^{n_l \times n_{l-1}}$
- Feature representation: $Z_l \in \mathbb{R}^{n_l \times N}$ is the output of layer l.
- Class means: Let μ_c be the mean of features in class c, and μ_G be the global mean.
- $s_1(A) \geq \cdots \geq s_K(A)$: singular values of A in non-increasing order
- $\kappa(A) = \dfrac{s_1(A)}{s_K(A)}$

Problem Setup

Network Architecture. Consider a deep neural network with L layers ($L = L_1 + L_2$),

- **Backbone (L_1 layers):** Non-linear layers with activation σ

$$Z_l = \sigma(W_l Z_{l-1}) \quad for\ l \leq L_1$$

- **Linear Head (L_2 layers):** Linear layers (no activation), assuming $L_2 \geq 2$

$$Z_l = W_l Z_{l-1} \quad for\ l > L_1$$

Metrics

NC1 (Within-Class Variability)

- We analyze the features at the penultimate layer Z_{L-1}.
- **Definition**: Ratio of Within-class covariance (Σ_W) to Between-class covariance (Σ_B).

$$NC1(Z_{L-1}) = \frac{\mathrm{tr}(\Sigma_W)}{\mathrm{tr}(\Sigma_B)}$$

  - $\Sigma_W = \frac{1}{N} \sum_{c,i} (z_{ci} - \mu_c)(z_{ci} - \mu_c)^\top$ (Distance to class mean)
  - $\Sigma_B = \frac{1}{K} \sum_{c=1}^{K} (\mu_c - \mu_G)(\mu_c - \mu_G)^\top$ (Distance between class means)

- **Goal**: Show that $NC1 \to 0$ (Collapse).

Metrics

NC2 (Simplex ETF)

- **Definition:** The condition number of the class-mean matrix of Z.

$$NC2(Z_{L-1}) = \kappa(\overline{Z}) = \frac{s_1(\overline{Z})}{s_K(\overline{Z})}\overline{Z}$$

  - $\overline{Z}$: The matrix of class means stacked as columns
  - $\kappa(\,\cdot\,)$: The condition number (ratio of the largest singular value to the smallest non-zero singular value).

- **Goal**: Show that $NC2 \rightarrow 1$ (Orthogonal).

Metrics

NC3 (Alignment)

- **Definition:** The average cosine similarity of features and weight vectors corresponding to the features' class.

$$NC3(Z_{L-1}) = \frac{1}{N} \sum_{c,i} \cos(z_{ci}, W_{c:})$$

  - z_{ci}: The feature vector of the i-th sample in class c.
  - $W_{c:}$: The row vector of the last layer's weight matrix W corresponding to class c.

- **Goal**: Show that $NC3 \rightarrow 1$ (Align).

Theoretical Picture

Part 1. The Geometric Conditions (Theorem 3.1)

- If Low Error (ϵ_1) & Balancedness (ϵ_2) → NC1 (Collapse).

Part 2. The Training Dynamics (Theorem 4.4)

- **Gradient Descent** minimizes error $(\epsilon_1 \to 0)$.
- **Weight Decay** aligns weights $(\epsilon_2 \to 0)$.

Overview

1. Preliminaries and Setup
2. **Part 1: Conditions for Neural Collapse**
3. Part 2: Gradient Descent Leads to NC
4. Numerical Results
5. Conclusion

Conditions for NC

Assumptions for Theorem 3.1.

If the network satisfies

- *approximate interpolation, i.e.,*

$$\|Z_L - Y\|_F \leq \epsilon_1$$

- *approximate balancedness, i.e.,*

$$\|W_{l+1}^{\top} W_{l+1} - W_l W_l^{\top}\|_{op} \leq \epsilon_2 \quad for \ l \in \{L_1 + 1, \ldots, L - 1\}$$

- *bounded representations and weights, i.e.,*

$$\|Z_{L-2}\|_{op}, \ \|Z_{L-1}\|_{op}, \ \|W_l\|_{op} \leq r \quad for \ l \in \{L_1 + 1, \ldots, L\}$$

Balancedness and Interpolation Imply Neural Collapse

Theorem 3.1. (partial) *Upper Bound on NC1.*

Then if $\epsilon_1 \leq \min\left(s_K(Y), \sqrt{\frac{(K-1)N}{4K}}\right)$,

$$NC1(Z_{L-1}) = \frac{\mathrm{tr}(\Sigma_W)}{\mathrm{tr}(\Sigma_B)}$$

Under these conditions, the Within-Class Variability collapses:

$$NC1(Z_{L-1}) = O((\epsilon_1 + \sqrt{\epsilon_2})^2)$$

Implication. As training error vanishes ($\epsilon_1 \to 0$) and layers become balanced ($\epsilon_2 \to 0$).

Consequently, $NC1(Z_{L-1}) \to 0$.

Proof Sketch

1. Key Idea: Orthogonal Decomposition

- To analyze the feature variability, we decompose Z_{L-1} based on the Row Space of the last weight W_L.

- Let P be the projection onto the Row Space of W_L.

$$Z_{L-1} = \underbrace{PZ_{L-1}}_{\text{Row space component}} + \underbrace{(I - P)Z_{L-1}}_{\text{Null space component}}$$

Proof Sketch

1. Key Idea: Orthogonal Decomposition

- To analyze the feature variability, we decompose Z_{L-1} based on the Row Space of the last weight W_L.
- Let P be the projection onto the Row Space of W_L.

$$Z_{L-1} = \underbrace{PZ_{L-1}}_{\text{Row space component}} + \underbrace{(I-P)Z_{L-1}}_{\text{Null space component}}$$

Controlled by

$\|Z_L - Y\|_F$

Interpolation

$\|W_{l+1}^{\top} W_{l+1} - W_l W_l^{\top}\|_{op}$

Balancedness

Bounding Each Component

(A) Row Space Component: Controlled by Interpolation (ϵ_1)

$$Z_{L-1} = PZ_{L-1} + (I - P)Z_{L-1} = W_L^+ Y + \boxed{W_L^+(Z_L - Y)} + (I - P)W_{L-1}Z_{L-2}.$$

- Since $W_L Z_{L-1} \approx Y$ (Low Error), this component is fixed to match the labels.

$$\boxed{\|W_L^+(Z_L - Y)\|_F} = O(\epsilon_1)$$

- W_L^+ : pseudoinverse of W_L

- $P = W_L^+ W_L$: projection on row space of W_L

- Since $s_K(W_L) \geq (s_K(Y) - \epsilon_1)/r$,

$$\boxed{\|W_L^+(Z_L - Y)\|_F} \leq \frac{\epsilon_1 r}{s_K(Y) - \epsilon_1}$$

Bounding Each Component

(B) Null Space Component: Controlled by Balancedness (ϵ_2)

$$Z_{L-1} = PZ_{L-1} + (I-P)Z_{L-1} = W_L^{+}Y + W_L^{+}(Z_L - Y) + (I-P)W_{L-1}Z_{L-2}.$$

- Usually, features can vary freely here (Noise).
- However, Balancedness implies layers align ($W_{L-1}W_{L-1}^{\top} \approx W_L^{\top}W_L$), suppressing this component.

$$\|(I-P)Z_{L-1}\|_F = O(\sqrt{\epsilon_2})$$

$$\begin{aligned}\|(I-P)W_{L-1}Z_{L-2}\|_F^2 &\leq \|(I-P)W_{L-1}\|_F^2\|Z_{L-2}\|_{op}^2 \leq r^2\mathrm{tr}((I-P)W_{L-1}W_{L-1}^{\top}) \\ &= r^2\mathrm{tr}((I-P)(W_{L-1}W_{L-1}^{\top} - W_L^{\top}W_L)) \leq r^2 n_{L-1}\epsilon_2\,.\end{aligned}$$

Proof Sketch

3. Conclusion

$$Z_{L-1} = PZ_{L-1} + (I-P)Z_{L-1} = W_L^+ Y + \cancelto{}{W_L^+(Z_L - Y)} + \cancelto{}{(I-P)W_{L-1}Z_{L-2}}$$

- Thus, Z_{L-1} collapses to a unique structure ($W_L^+ Y$), implying NC1.

$$\|Z_{L-1} - W_L^+ Y\|_F \leq r \left(\frac{\epsilon_1}{s_K(Y) - \epsilon_1} + \sqrt{n_{L-1}\epsilon_2} \right)$$

Finally,

- $\mathrm{tr}(\Sigma_W) \leq \dfrac{1}{N} \|Z_{L-1} - W_L^+ Y\|_F^2$
- $\mathrm{tr}(\Sigma_B) = \Omega(1)$
- $NC1(Z_{L-1}) = \dfrac{\mathrm{tr}(\Sigma_W)}{\mathrm{tr}(\Sigma_B)} = O((\epsilon_1 + \sqrt{\epsilon_2})^2)$

Overview

1. Preliminaries and Setup
2. Part 1: Conditions for Neural Collapse
3. **Part 2: Gradient Descent Leads to NC**
4. Numerical Results
5. Conclusion

Training Setup

- **λ-regularized square loss**
  - $C_{\lambda}(\theta) = \frac{1}{2}\|Z_L(\theta) - Y\|_F^2 + \frac{\lambda}{2}\|\theta\|_2^2$
- **GD update**
  - $\theta_{k+1} = \theta_k - \eta \nabla C_{\lambda}(\theta_k)$,
- **Notations**
  - step size η
  - $\theta_k = (W_l^k)_{=1}^L$ contains all parameters at step k
  - Z_l^k denotes the output of layer l after k steps

Assumptions

Assumption 4.1. *(Pyramidal network topology) Let $n_1 \geq N$ and $n_2 \geq n_3 \geq \ldots \geq n_L$.*

Assumption 4.2. *(Activation function)*

Fix $\gamma \in (0,1)$ and $\beta \geq 1$. Let σ satisfy that:

(i) $\sigma'(x) \in [\gamma, 1]$,

(ii) $|\sigma(x)| \leq |x|$ *for every* $x \in \mathbb{R}$,

(iii) σ' *is* β*-Lipschitz.*

- Smooth leaky ReLUs satisfies **Assumption 4.2**, and they uniformly approximate the ReLU over $\mathbb{R}$. [2]

[2] Nguyen & Mondelli, Global convergence of deep networks with one wide layer followed by pyramidal topology, NeurIPS 2020.

Assumptions

Notations. At initialization $\theta_0 = (W_l^0)_{l=0}^{L}$,

- $\lambda_\ell = \sigma_{\min}(W_l^0), \quad \bar{\lambda}_l = \|W_l^0\|_{op} + \min_{l \in \{3,\ldots,L\}} \lambda_l$
- $\lambda_{i \to j} = \prod_{l=i}^{j} \lambda_l, \quad \bar{\lambda}_{i \to j} = \prod_{l=i}^{j} \bar{\lambda}_l$
- $\lambda_F = \sigma_{\min}\left(\sigma(W_0^1 X)\right)$

Assumption 4.3. *(Initial conditions)*

$$\lambda_F \lambda_{3 \to L} \min\left(\lambda_F, \min_{l \in \{3,\ldots,L\}} \lambda_l\right) \geq 8\gamma \sqrt{\left(\frac{2}{\gamma}\right)^L C_0(\theta_0)} \,.$$

Gradient Descent Leads to NC1

Theorem 4.4. *Let the network satisfy* **Assumptions 4.1, 4.2***, and* **4.3.** *Fix* $0 < \epsilon_1 \leq \frac{1}{2}\sqrt{\frac{(K-1)N}{K}}$, $\epsilon_2 > 0$*, let* $b \geq 1$ *be s.t.* $\|X_{:i}\|_2 \leq b$ *for all* i*, and run* k *steps of* λ*-regularized GD with step size* η*, where* $\lambda = \Theta(\epsilon_1^2)$*,* $\eta = O(\epsilon_2)$ *and* $k = \Omega\left(\frac{1}{\epsilon_1^2\epsilon_2}\log\frac{1}{\epsilon_2}\right)$*. Then,*

$$NC1(Z_{L-1}^k) = O\left((\epsilon_1 + \sqrt{\epsilon_2})^2\right).$$

- Sufficiently small **regularization** & **learning rate**
- Sufficiently long **GD**

⇒ **Within-class variability vanishes**

Proof Sketch

- We will show that λ-regularized GD fulfills the conditions for NC1 (Theorem 3.1)
    - (Interpolation) $\|Z_L - Y\|_F \leq \epsilon_1$
    - (Balancedness) $\|W_{l+1}^\top W_{l+1} - W_l W_l^\top\|_{op} \leq \epsilon_2$

- Distinguish two phases in the training dynamics
    - **First phase.** The loss $\|Z_L - Y\|_F$ decreases exponentially
    - **Second phase.** $\|W_{l+1}^\top W_{l+1} - W_l W_l^\top\|_{op}$ decreases exponentially at a rate $1/\lambda$

Interpolation

$$C_\lambda(\theta) = \frac{1}{2}\|Z_L - Y\|_F^2 + \frac{\lambda}{2}\|\theta\|_2^2$$

We will show

- $C(\theta)$ satisfies the α-**Polyak-Lojasiewicz (PL)** inequality
- $\nabla C(\theta)$ is β**-Lipschitz** ($C(\theta)$ β-smoothness)

Then,

- $C(\theta_k) - C^* \leq (C(\theta_0) - C^*)\left(1 - \dfrac{\alpha}{\beta}\right)^k$
- $C(\theta_{k_1}) \leq \epsilon_1^2$, for some step k_1

PL condition

- By Lemma in [1], unregularized loss $C_0(\theta)$ satisfies the α-**PL inequality,**

$$\|\nabla C_0(\theta)\|_2^2 \geq \frac{\alpha}{2} C_0(\theta),$$

for all θ in a ball $B(\theta_0, r_0)$ centered at initialization with a sufficiently large radius r_0.

Proposition 4.5. (partial) *Let $C_0(\theta)$ satisfy the α-PL inequality in the ball $B(\theta_0, r_0)$. Then, in the same ball, $C_\lambda(\theta)$ satisfies the inequality*

$$\|\nabla C_\lambda(\theta)\|_2^2 \geq \frac{\alpha}{4}(C_\lambda(\theta) - \lambda m_\lambda),$$

where $m_\lambda = (1 + \sqrt{4\lambda/\alpha})^2(\|\theta_0\|_2 + r_0)^2$.

[2] Nguyen & Mondelli, Global convergence of deep networks with one wide layer followed by pyramidal topology, NeurIPS 2020.

PL condition

- **Proof.** Let $\theta \in B(\theta_0, r_0)$, then

$$
\begin{aligned}
\|\nabla C_0(\theta) + \lambda\theta\|_2^2 &\ge (\|\nabla C_0(\theta)\|_2 - \lambda\|\theta\|_2)^2 \\
&\ge \left(\sqrt{\frac{\alpha}{2}C_0(\theta)} - \lambda\|\theta\|_2\right)^2 \qquad \|\nabla C_0(\theta)\|_2^2 \ge \frac{\alpha}{2}C_0(\theta) \\
&= \left(\sqrt{\frac{\alpha}{2}C_\lambda(\theta) - \frac{\alpha\lambda}{4}\|\theta\|_2^2} - \lambda\|\theta\|_2\right)^2 \qquad \sqrt{a-b} \ge \sqrt{a} - \sqrt{b} \\
&\ge \left(\sqrt{\frac{\alpha}{2}C_\lambda(\theta)} - \left(\lambda + \sqrt{\frac{\alpha\lambda}{4}}\right)\|\theta\|_2\right)^2 \\
&\ge \frac{\alpha}{4}C_\lambda(\theta) - \lambda\left(\sqrt{\frac{\alpha}{4}} + \sqrt{\lambda}\right)^2\|\theta\|_2^2 \qquad (a-b)^2 \ge \frac{a^2}{2} - b^2 \\
&\ge \frac{\alpha}{4}\left[C_\lambda(\theta) - \lambda\left(1 + \sqrt{\frac{4\lambda}{\alpha}}\right)^2(\|\theta_0\|_2 + r_0)^2\right].
\end{aligned}
$$

Smoothness

Lemma C.1. *Let* $b \geq 1$ *be s.t.* $\|X_{:i}\|^2 \leq b$ *for all* i*. and* $\|W_l\|_{op} \leq r_l$ *for all* $l \in [L]$ *and* $r_l \geq 1$*.*

Then the gradient $\nabla C_0(\theta)$ *is* $5N\beta b^3 \left(\prod_{j=1}^{L} r_j \right)^3 L^{5/2}$*-Lipschitz.*

Summing up

Proposition 4.5. (cont'd) *Assume that* $r_0 \geq 8\sqrt{C_\lambda(\theta_0)/\alpha}$ *and* $\nabla C_0(\theta)$ *is* β_1*-Lipschitz in* $B(\theta_0, r_0)$*.*

Then, for any $\eta < 1/(2\beta_1)$*, there exists*

$$k_1 \leq \left\lceil \frac{\log \frac{\lambda m_\lambda}{C_\lambda(\theta_0) - \lambda m_\lambda}}{\log(1 - \eta\frac{\alpha}{8})} \right\rceil$$

such that the k_1*-th iterate of GD satisfies*

$$C_\lambda(\theta_{k_1}) \leq 2\lambda m_\lambda, \qquad \|\theta_{k_1} - \theta_0\|_2 \leq 8\sqrt{\frac{C_\lambda(\theta_0)}{\alpha}} \leq r_0\,.$$

- By upper bounding λ, we can bound $C_\lambda(\theta_{k_1}) \leq 2\lambda m_\lambda \leq \epsilon_1^2$
- Then the network achieves approximate interpolation, i.e.

$$\|Z_L - Y\|_F \leq \sqrt{2C_\lambda(\theta_{k+1})} \leq \epsilon_1\sqrt{2}$$

Proof

$$\begin{aligned}
C_\lambda(\theta_{k+1}) - C_\lambda(\theta_k) &= -\eta \int_0^1 \langle \nabla C_\lambda(\theta_k - s\eta \nabla C_\lambda(\theta_k)), \nabla C_\lambda(\theta_k) \rangle ds \\
&= -\eta \|\nabla C_\lambda(\theta_k)\|_2^2 + \eta \int_0^1 \langle \nabla C_\lambda(\theta_k - s\eta \nabla C_\lambda(\theta_k)) - \nabla C_\lambda(\theta_k), \nabla C_\lambda(\theta_k) \rangle ds \\
&\leq -\eta \|\nabla C_\lambda(\theta_k)\|_2^2 + \eta \int_0^1 \|\nabla C_\lambda(\theta_k)\|_2 \|\nabla C_\lambda(\theta_k - s\eta \nabla C_\lambda(\theta_k)) - \nabla C_\lambda(\theta_k)\|_2 ds \\
&\leq -\eta \|\nabla C_\lambda(\theta_k)\|_2^2 + \eta^2 \beta_1 \|\nabla C_\lambda(\theta_k)\|_2^2 \\
&= -\eta(1 - \eta\beta_1)\|\nabla C_\lambda(\theta_k)\|_2^2 \\
&\leq -\frac{\eta}{2}\|\nabla C_\lambda(\theta_k)\|_2^2 \\
&\leq -\eta \frac{\alpha}{8}(C_\lambda(\theta_k) - \lambda m_\lambda)\,.
\end{aligned}$$

Similar to the process in the lecture…

Post-convergence

- **Phase 1** ($k < k_1$).
  - Satisfying PL-condition
  - $\theta \in B(\theta_0, r_0)$, for $r_0 \geq 8\sqrt{C_\lambda(\theta_0)/\alpha}$
- **Phase 2** ($k \geq k_1$).
  - θ might not be in $B(\theta_0, r_0)$, but satisfying $C_\lambda \leq \epsilon_1^2$.

$$\frac{\lambda}{2}\|\theta\|^2 \leq C_\lambda(\theta) \leq \epsilon_1^2 \implies \|\theta\| \leq \epsilon_1\sqrt{\frac{2}{\lambda}}$$

  - Need new β_2-smoothness, based on this bound
  - Set $\eta \leq \min\left(1/(2\beta_1), 1/\beta_2\right)$, guaranteeing that the loss decreases globally

Proof Sketch

- We have to show that λ-regularized GD fulfills the conditions for NC1 (Theorem 3.1)
  - (Interpolation) $\|Z_L - Y\|_F \leq \epsilon_1$
  - (Balancedness) $\|W_{l+1}^\top W_{l+1} - W_l W_l^\top\|_{op} \leq \epsilon_2$
- Distinguish two phases in the training dynamics
  - **First phase.** The loss $\|Z_L - Y\|_F$ decreases exponentially
  - **Second phase.** $\|W_{l+1}^\top W_{l+1} - W_l W_l^\top\|_{op}$ decreases exponentially at a rate $1/\lambda$

Balancedness

- **Objective**. Bounding $\|W_{l+1}^\top W_{l+1} - W_l W_l^\top\|$
- **Define.**
  - $D_l^{(k)} = W_{l+1}^\top W_{l+1} - W_l W_l^\top$
  - T_l^k : Gradient component except weight decay
- **Update rule**

$$W_l^{k+1} = (1-\eta\lambda) W_l^k - \eta T_l^k$$

$$D_l^{k+1} = (1-\eta\lambda)^2 D_l^k + \eta^2((T_{l+1}^k)^\top T_{l+1}^k - T_l^k (T_l^k)^\top)$$

  - Since $(1-\eta\lambda)^2 < 1$, it forces D_l to shrink over time exponentially

Balancedness

- **Upper bounding the noise term**
  - After phase 1 training, $\|Z_L^k - Y\|_F \leq \epsilon_1\sqrt{2}$. T_l^k is also bounded and very small.
  - Substituting these bounds yields the upper bound for the noise term:

$$\|D_l^{k+1}\|_{op} \leq (1-\eta\lambda)^2\|D_l^k\|_{op} + 4\eta^2\epsilon_1^2\left(\frac{2\epsilon_1^2}{\lambda}\right)^{L_1+L-1}\|X\|_{op}^2$$

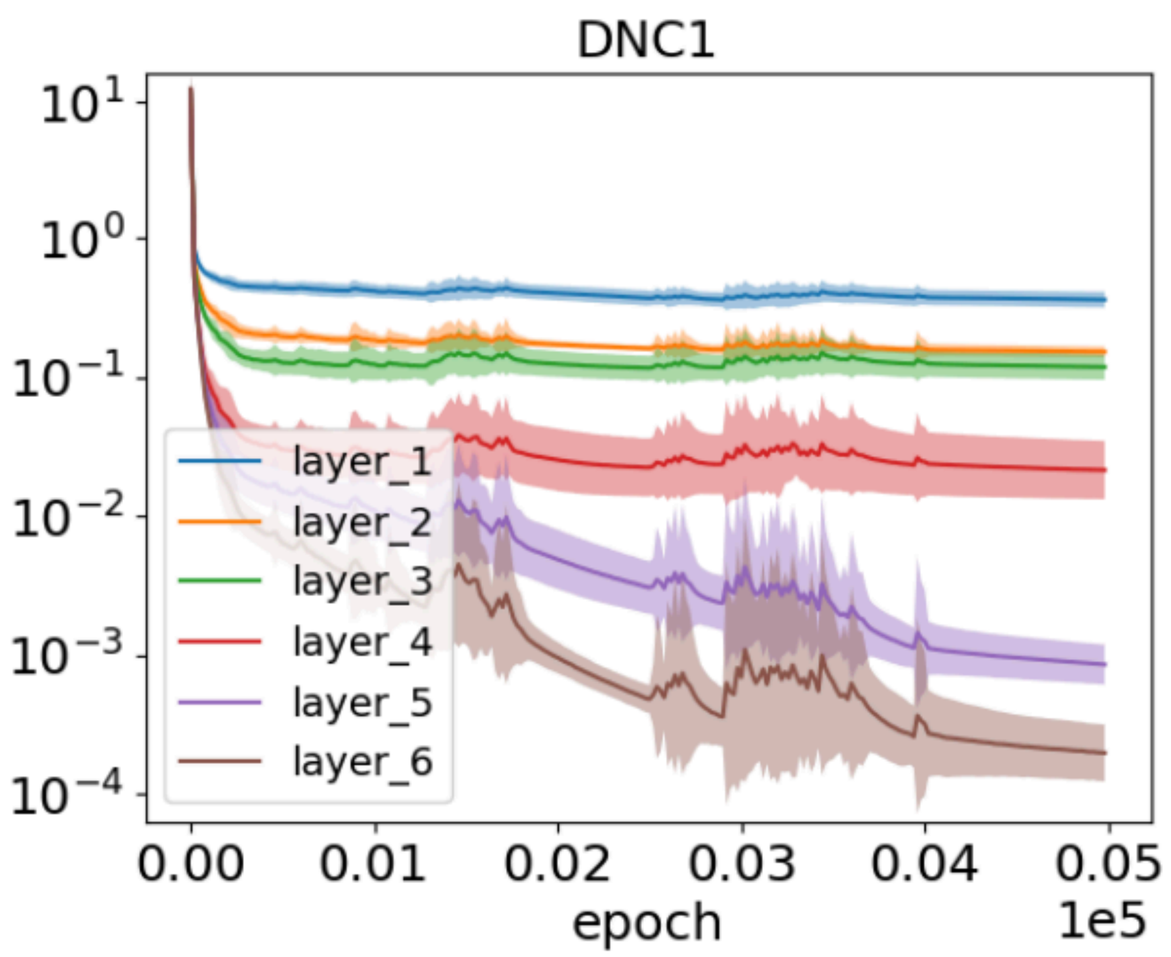
- **Convergence**
  - Selecting proper η and λ
    - if $\|D_l^k\|_{op} > \epsilon_2$, then $\|D_l^{k+1}\|_{op} \leq (1-\eta\lambda)\|D_l^k\|_{op}$;
    - if $\|D_l^k\|_{op} \leq \epsilon_2$, then $\|D_l^{k+1}\|_{op} \leq (1-\eta\lambda)\epsilon_2 \leq \epsilon_2$.

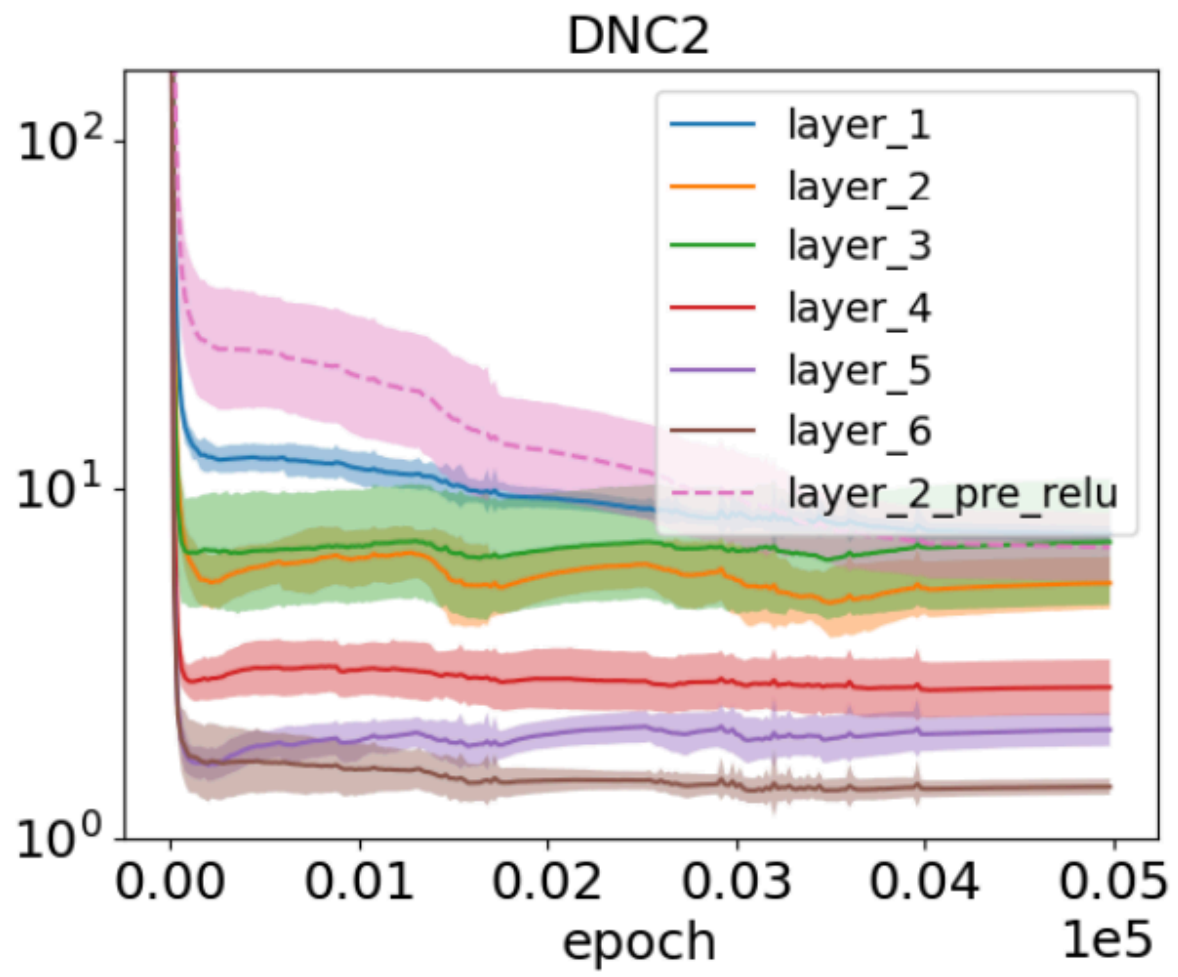
Overview

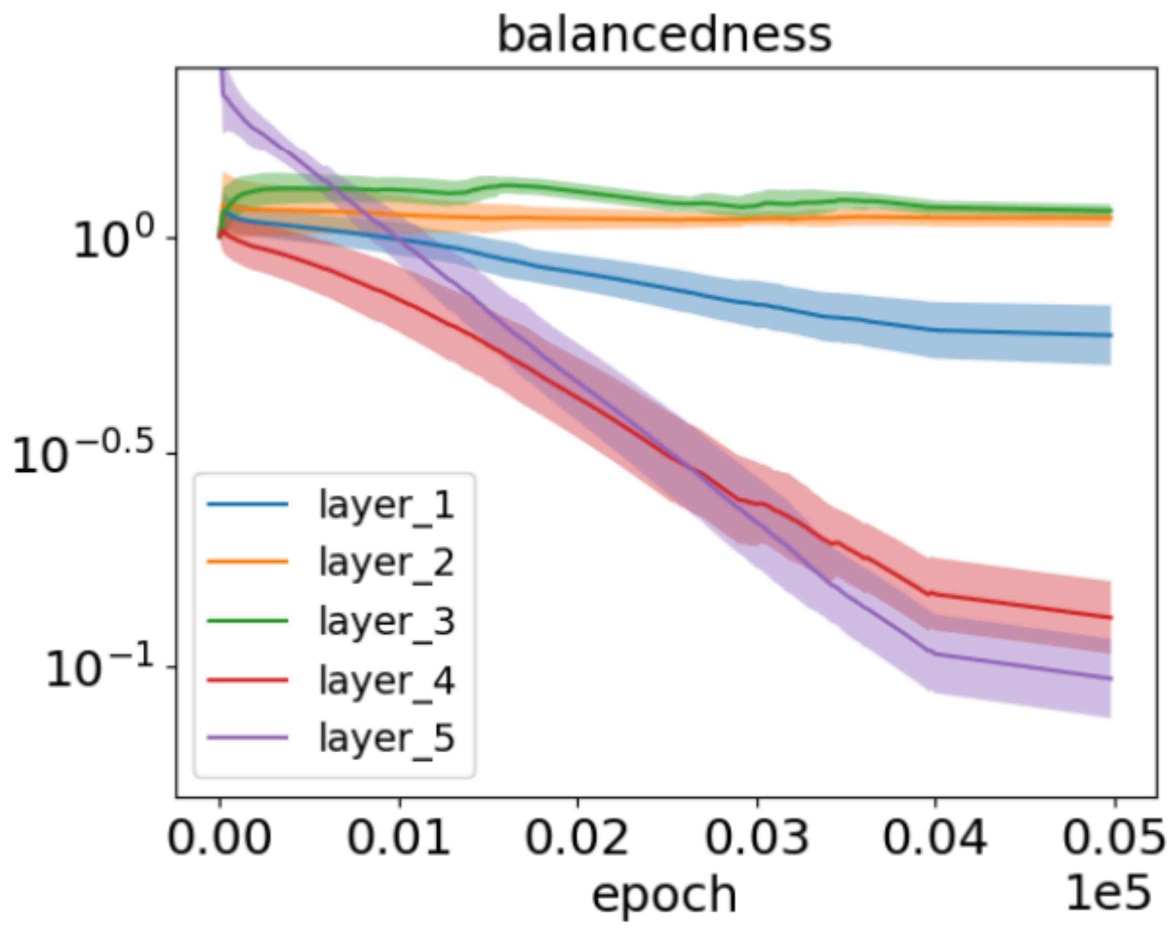
1. Preliminaries and Setup
2. Part 1: Conditions for Neural Collapse
3. Part 2: Gradient Descent Leads to NC
4. **Numerical Results**
5. Conclusion

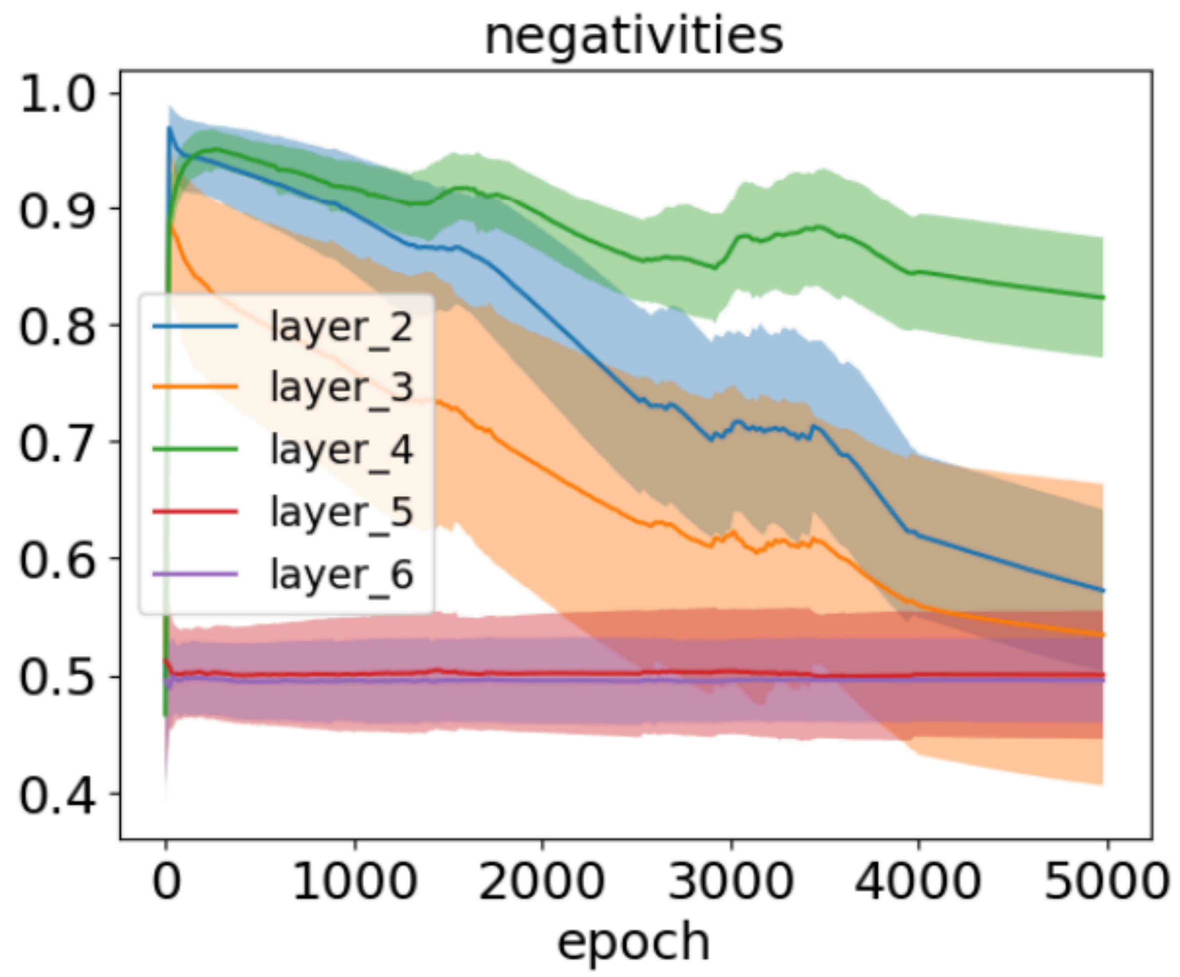
NC metric and Balancedness

- **Setup**. ResNet20 + (6 linear + 3 ReLU) / CIFAR10
- **Metric.** NC1, NC2, Balancedness $\frac{\left\| W_{\ell+1}^{\top} W_{\ell+1} - W_\ell W_\ell^{\top} \right\|_{op}}{\min\left\{ \left\| W_{\ell+1}^{\top} W_{\ell+1} \right\|_{op}, \left\| W_\ell W_\ell^{\top} \right\|_{op} \right\}}$, negativity $\frac{\|Z_\ell - \sigma(Z_\ell)\|_{op}}{\|Z_\ell\|_{op}}$.
- **Results**. NC1 and Balancedness converge exponentially during training



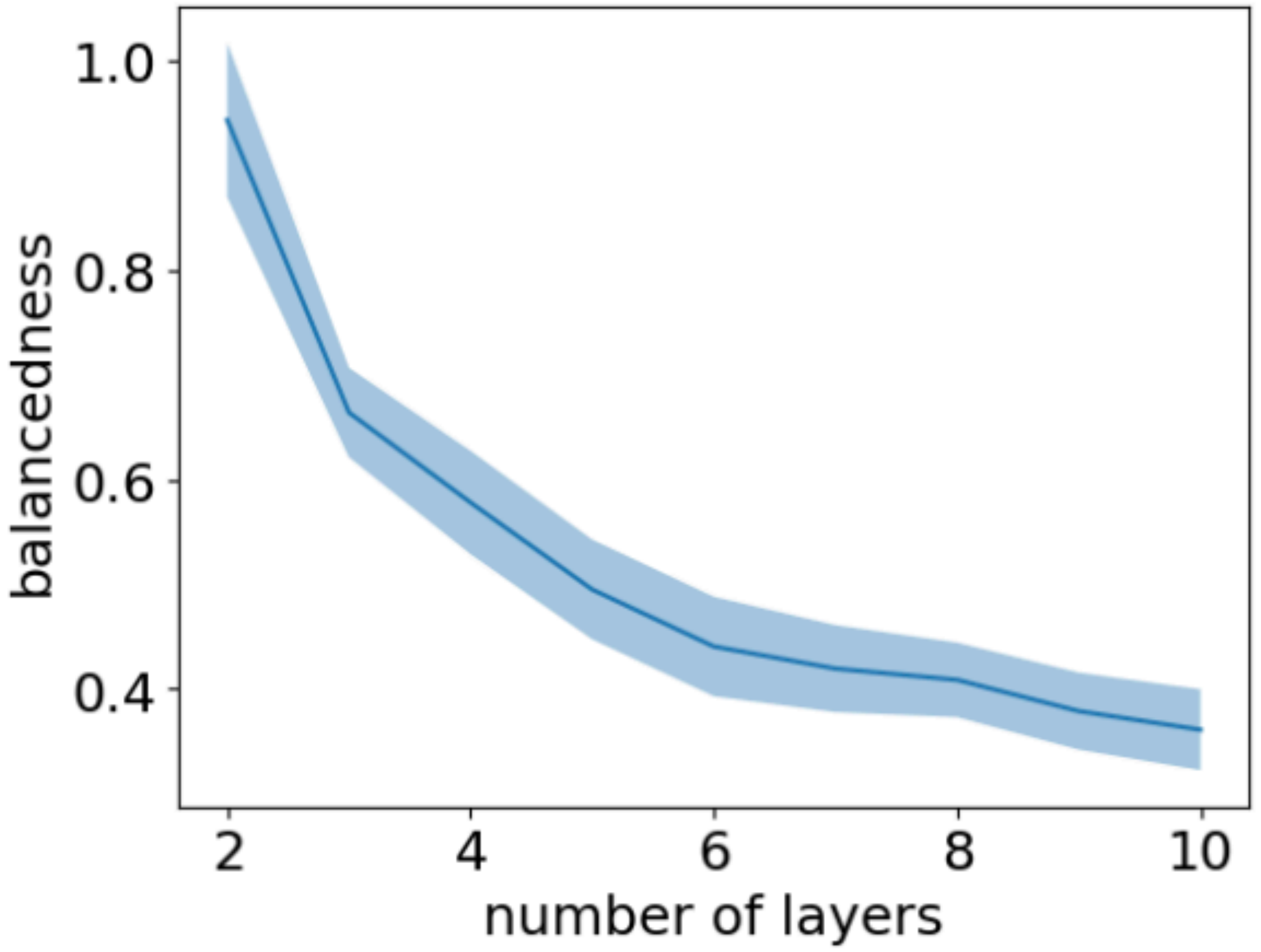


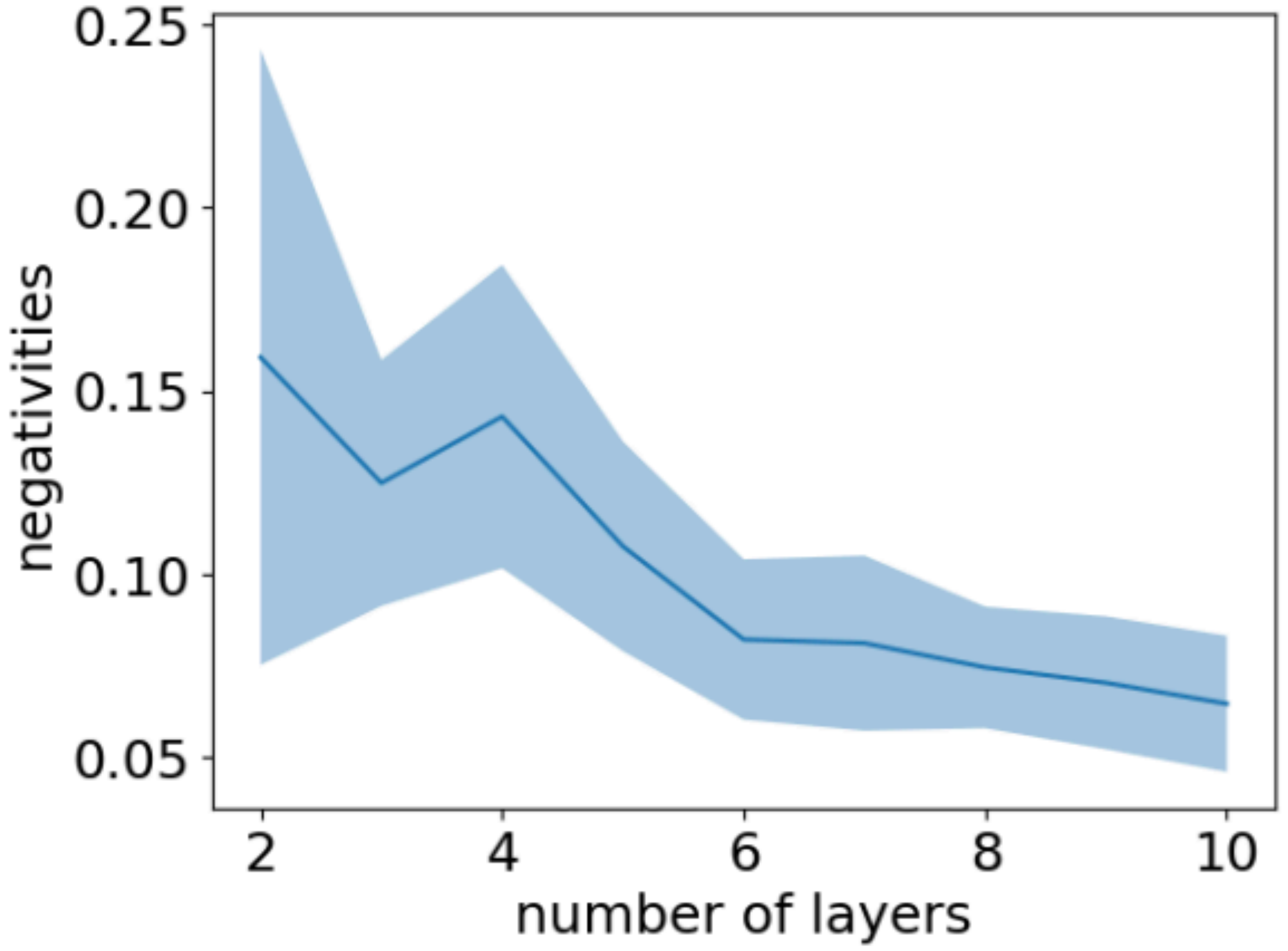




Non-linear layers are balanced, linear

- **X axis**. Depth of the non-linear part
- **Y axis**. balancedness, negativity (non-linearity)
- **Results**. Non-linear layers are increasingly balanced and linear, as the depth of the non-linear part increases.





Overview

1. Preliminaries and Setup
2. Part 1: Conditions for Neural Collapse
3. Part 2: Gradient Descent Leads to NC
4. Numerical Results
5. **Conclusion**

Conclusion

Summary

- Weight Decay + GD → Balancedness & Interpolation → NC1 (Collapse)
- Deep Linear Head → Good Conditioning → NC2 (Orthogonality) & NC3 (Alignment)

Key Insight

- ***Weight decay*** induces ***balancedness*** between adjacent linear layers
- This balancedness suppresses the unnecessary variance (null-space) of the features, serving as the crucial link that eventually leads to NC1

Limitation

- The theoretical proofs rely on the assumption that the linear head consists of ***at least two linear layers***

Q&A

Reference

1. Papyan et al., Prevalence of neural collapse during the terminal phase of deep learning training, PNAS 2020.

2. Nguyen & Mondelli, Global convergence of deep networks with one wide layer followed by pyramidal topology, NeurIPS 2020.