Simplex symmetry in the final and penultimate layers of neural network classifiers

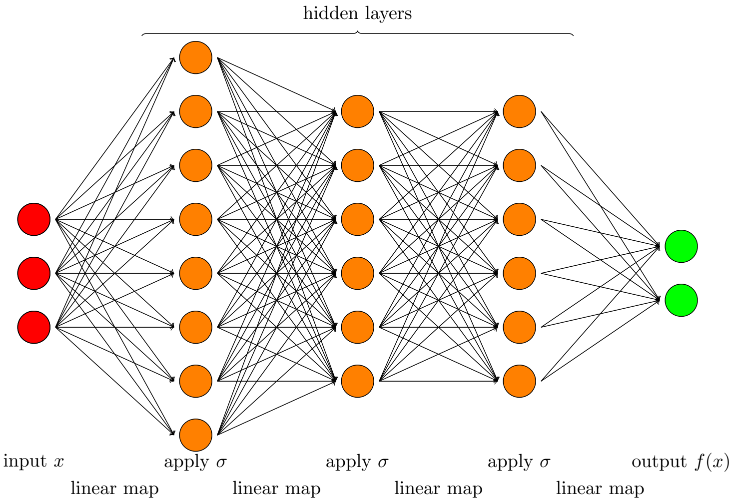
Stephan Wojtowytsch
joint work with Weinan E

Princeton University

July 28, 2021

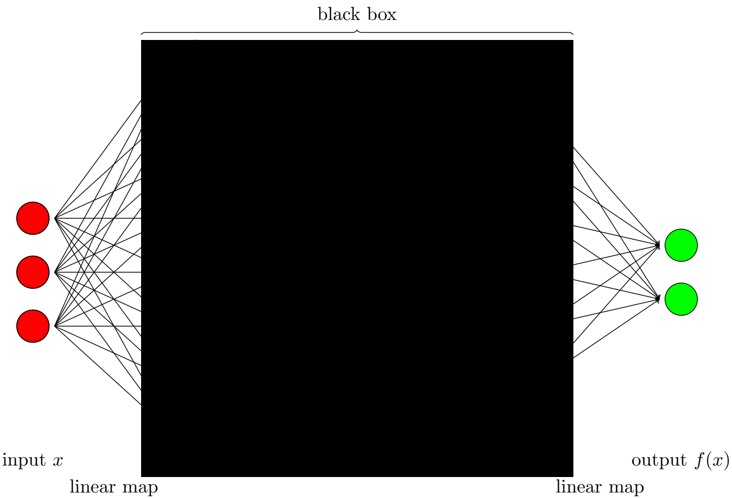
MSML 2021

Deep Neural networks

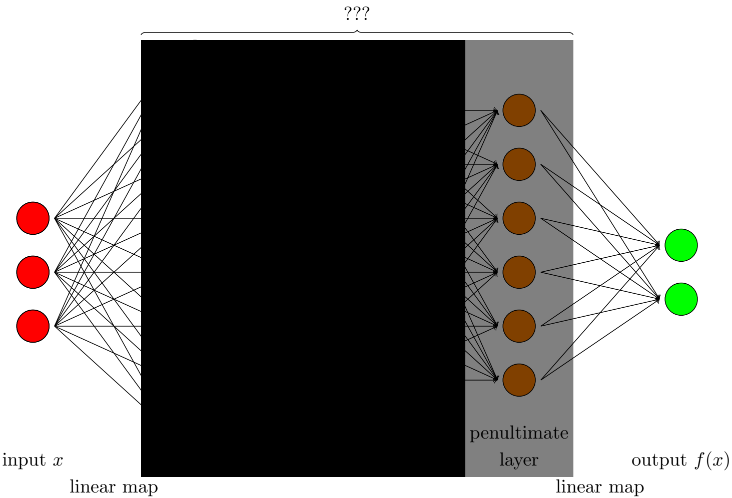


Deep Neural networks

PRINCETON UNIVERSITY



Deep Neural networks



RESEARCH ARTICLE

Prevalence of neural collapse during the terminal phase of deep learning training

Vardan Papyan, X. Y. Han, and David L. Donoho

PNAS October 6, 2020 117 (40) 24652-24663; first published September 21, 2020;
https://doi.org/10.1073/pnas.2015509117

Contributed by David L. Donoho, August 18, 2020 (sent for review July 22, 2020; reviewed by Helmut Boelsckei and Stéphane Mallat)

See related content:

Another step toward demystifying deep neural networks - Oct 15, 2020

Numerical observations

Consider a hypothesis class of functions of the form $h(x) = A\,f(x)$ and risk functionals

$$\widehat{\mathcal{R}}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell_{ce}\big(h(x_i), c_i\big) \qquad \text{or} \qquad \mathcal{R}(h) = \mathbb{E}_{(x,c)\sim\mathbb{P}}\big[\ell_{ce}\big(h(x), c\big)\big]$$

where

$$\ell_{ce}(h, c) = -\log\left(\frac{\exp(h_c)}{\sum_j \exp(h_j)}\right)$$

is the cross entropy loss function.

Numerical observations

Consider a hypothesis class of functions of the form $h(x) = A\,f(x)$ and risk functionals

$$\widehat{\mathcal{R}}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell_{ce}\big(h(x_i), c_i\big) \qquad \text{or} \qquad \mathcal{R}(h) = \mathbb{E}_{(x,c)\sim\mathbb{P}}\big[\ell_{ce}\big(h(x), c\big)\big]$$

where

$$\ell_{ce}(h, c) = -\log\left(\frac{\exp(h_c)}{\sum_j \exp(h_j)}\right)$$

is the cross entropy loss function. For deep neural network classifiers trained by SGD with cross-entropy loss, the following become asymptotically true:

1. f maps all points in the class C_i to a single value y_i.

Numerical observations

Consider a hypothesis class of functions of the form $h(x) = A\,f(x)$ and risk functionals

$$\widehat{\mathcal{R}}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell_{ce}\big(h(x_i), c_i\big) \qquad \text{or} \qquad \mathcal{R}(h) = \mathbb{E}_{(x,c)\sim\mathbb{P}}\big[\ell_{ce}\big(h(x), c\big)\big]$$

where

$$\ell_{ce}(h, c) = -\log\left(\frac{\exp(h_c)}{\sum_j \exp(h_j)}\right)$$

is the cross entropy loss function. For deep neural network classifiers trained by SGD with cross-entropy loss, the following become asymptotically true:

1. f maps all points in the class C_i to a single value y_i.
2. the distance $\|y_i - y_j\|$ and scalar product $\langle y_i - y_l, y_l - y_j\rangle$ are independent of i, j, l, i.e. y_i are the vertices of a regular $k-1$-dimensional standard simplex in $\mathbb{R}^m$.

Numerical observations

Consider a hypothesis class of functions of the form $h(x) = A\,f(x)$ and risk functionals

$$\widehat{\mathcal{R}}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell_{ce}\big(h(x_i), c_i\big) \quad \text{or} \quad \mathcal{R}(h) = \mathbb{E}_{(x,c)\sim\mathbb{P}}\big[\ell_{ce}\big(h(x), c\big)\big]$$

where

$$\ell_{ce}(h, c) = -\log\left(\frac{\exp(h_c)}{\sum_j \exp(h_j)}\right)$$

is the cross entropy loss function. For deep neural network classifiers trained by SGD with cross-entropy loss, the following become asymptotically true:

1. f maps all points in the class C_i to a single value y_i.
2. the distance $\|y_i - y_j\|$ and scalar product $\langle y_i - y_l, y_l - y_j\rangle$ are independent of i, j, l, i.e. y_i are the vertices of a regular $k-1$-dimensional standard simplex in $\mathbb{R}^m$.
3. the i-th row of A is parallel to $y_i - \frac{1}{k}\sum_j y_j$.

Optimality of neural collapse

Final layer geometry: collapse to a point

Lemma

Let $h \in \mathcal{H}$ and set

$$z_i := \frac{1}{|C_i|} \int_{C_i} h(x')\, \mathbb{P}(\mathrm{d}x'), \qquad \bar{h}(x) = z_i \quad \text{for all } x \in C_i.$$

Then $\mathcal{R}(\bar{h}) \leq \mathcal{R}(h)$.

Proof.

Jensen's inequality. $\square$

PRINCETON UNIVERSITY

Lemma

Let $h \in \mathcal{H}$ and set

$$z_i := \frac{1}{|C_i|}\int_{C_i} h(x')\,\mathbb{P}(\mathrm{d}x'), \qquad \bar{h}(x) = z_i \quad \text{for all } x \in C_i.$$

Then $\mathcal{R}(\bar{h}) \leq \mathcal{R}(h)$.

Proof.

Jensen's inequality. □

Corollary

If $\mathcal{H}$ is the class of $\mathbb{P}$-measurable functions from $\mathbb{R}^d$ into a convex compact set $V \subset \mathbb{R}^k$, then any minimizer h of $\mathcal{R}$ in $\mathcal{H}$ maps the class C_i to a single point $z_i \in V$ for all $i = 1, \dots, k$.

PRINCETON UNIVERSITY

Lemma (E-W '20)

Let $B_R(0)$ be the ball of radius $R > 0$ in $\mathbb{R}^k$ with respect to the ℓ^p-norm, $1 < p < \infty$. For every i there exists a unique minimizer z_i of

$$\Phi_i(z) = -\log\left(\frac{\exp(z\cdot e_i)}{\sum_{j=1}^k \exp(z\cdot e_j)}\right)$$

in $B_R(0)$ and $z_i = R\left(\alpha\, e_i + \beta \sum_{j\neq i} e_j\right)$ for $\alpha, \beta \in \mathbb{R}$ which only depend on p.

Lemma (E-W '20)

Let $B_R(0)$ be the ball of radius $R > 0$ in $\mathbb{R}^k$ with respect to the ℓ^p-norm, $1 < p < \infty$. For every i there exists a unique minimizer z_i of

$$\Phi_i(z) = -\log\left(\frac{\exp(z \cdot e_i)}{\sum_{j=1}^k \exp(z \cdot e_j)}\right)$$

in $B_R(0)$ and $z_i = R\left(\alpha\, e_i + \beta \sum_{j \neq i} e_j\right)$ for $\alpha, \beta \in \mathbb{R}$ which only depend on p.

Corollary (E-W '20)

If $\mathcal{H}$ is the hypothesis class of $\mathbb{P}$-measurable functions from $\mathbb{R}^d$ to the ℓ^p-ball of radius $R > 0$, the unique minimizer h of $\mathcal{R}$ in $\mathcal{H}$ maps all $x \in C_i$ to z_i.

Penultimate layer optimality

Corollary (E-W '20)

For any $m \geq k-1$, consider the hypothesis class

$$\mathcal{H} = \left\{ h : \mathbb{R}^d \to \mathbb{R}^k \;\middle|\; h = Af \text{ where } \begin{array}{ll} f : \mathbb{R}^d \to \mathbb{R}^m \text{ is } \mathbb{P}-\text{measurable}, & \|f(x)\|_{\ell^2} \leq R \text{ a.e.} \\ A : \mathbb{R}^m \to \mathbb{R}^k \text{ is linear}, & \|A\|_{L(\ell^2,\ell^2)} \leq 1 \end{array} \right\}.$$

Penultimate layer optimality

Corollary (E-W '20)

For any $m \geq k-1$, consider the hypothesis class

$$\mathcal{H} = \left\{ h : \mathbb{R}^d \to \mathbb{R}^k \;\middle|\; h = Af \text{ where } \begin{array}{ll} f : \mathbb{R}^d \to \mathbb{R}^m \text{ is } \mathbb{P}-\text{measurable}, & \|f(x)\|_{\ell^2} \leq R \text{ a.e.} \\ A : \mathbb{R}^m \to \mathbb{R}^k \text{ is linear}, & \|A\|_{L(\ell^2,\ell^2)} \leq 1 \end{array} \right\}.$$

Then the unique minimizer $h \in \mathcal{H}$ of $\mathcal{R}$ satisfies $h = Af$ where

1. *there exist values $y_i \in \mathbb{R}^m$ such that $f(x) = y_i$ for almost every $x \in C_i$,*

Penultimate layer optimality

Corollary (E-W '20)

For any $m \geq k-1$, consider the hypothesis class

$$\mathcal{H} = \left\{ h : \mathbb{R}^d \to \mathbb{R}^k \;\middle|\; h = Af \text{ where } \begin{array}{ll} f : \mathbb{R}^d \to \mathbb{R}^m \text{ is } \mathbb{P}-\text{measurable}, & \|f(x)\|_{\ell^2} \leq R \text{ a.e.} \\ A : \mathbb{R}^m \to \mathbb{R}^k \text{ is linear}, & \|A\|_{L(\ell^2,\ell^2)} \leq 1 \end{array} \right\}.$$

Then the unique minimizer $h \in \mathcal{H}$ of $\mathcal{R}$ satisfies $h = Af$ where

1. *there exist values $y_i \in \mathbb{R}^m$ such that $f(x) = y_i$ for almost every $x \in C_i$,*
2. *the points y_i form the vertices of a regular $k-1$-dimensional simplex in $\mathbb{R}^m$,*

Corollary (E-W '20)

For any $m \geq k-1$, consider the hypothesis class

$$\mathcal{H} = \left\{ h : \mathbb{R}^d \to \mathbb{R}^k \,\middle|\, h = Af \text{ where } \begin{array}{ll} f : \mathbb{R}^d \to \mathbb{R}^m \text{ is } \mathbb{P}-\text{measurable}, & \|f(x)\|_{\ell^2} \leq R \text{ a.e.} \\ A : \mathbb{R}^m \to \mathbb{R}^k \text{ is linear}, & \|A\|_{L(\ell^2,\ell^2)} \leq 1 \end{array} \right\}.$$

Then the unique minimizer $h \in \mathcal{H}$ of $\mathcal{R}$ satisfies $h = Af$ where

1. *there exist values $y_i \in \mathbb{R}^m$ such that $f(x) = y_i$ for almost every $x \in C_i$,*
2. *the points y_i form the vertices of a regular $k-1$-dimensional simplex in $\mathbb{R}^m$,*
3. *the center of mass of the points y_i (with respect to the uniform distribution) is at the origin, and*

PRINCETON UNIVERSITY

Corollary (E-W '20)

For any $m \geq k-1$, consider the hypothesis class

$$\mathcal{H} = \left\{ h : \mathbb{R}^d \to \mathbb{R}^k \;\middle|\; h = Af \text{ where } \begin{array}{ll} f : \mathbb{R}^d \to \mathbb{R}^m \text{ is } \mathbb{P}-\text{measurable}, & \|f(x)\|_{\ell^2} \leq R \text{ a.e.} \\ A : \mathbb{R}^m \to \mathbb{R}^k \text{ is linear}, & \|A\|_{L(\ell^2,\ell^2)} \leq 1 \end{array} \right\}.$$

Then the unique minimizer $h \in \mathcal{H}$ of $\mathcal{R}$ satisfies $h = Af$ where

1. *there exist values $y_i \in \mathbb{R}^m$ such that $f(x) = y_i$ for almost every $x \in C_i$,*
2. *the points y_i form the vertices of a regular $k-1$-dimensional simplex in $\mathbb{R}^m$,*
3. *the center of mass of the points y_i (with respect to the uniform distribution) is at the origin, and*
4. *A is an isometric embedding of the $k-1$-dimensional space spanned by $\{y_1, \ldots, y_k\}$ into $\mathbb{R}^k$.*

Interpretation

1. If $\mathcal{H}$ is very expressive, it is best to collapse all data points to the vertices of a standard simplex.

Interpretation

1. If $\mathcal{H}$ is very expressive, it is best to collapse all data points to the vertices of a standard simplex. This argument is not related to optimization algorithms, but the behavior is expected in the long term limit, as we approach minimal configurations.

Interpretation

1. If $\mathcal{H}$ is very expressive, it is best to collapse all data points to the vertices of a standard simplex. This argument is not related to optimization algorithms, but the behavior is expected in the long term limit, as we approach minimal configurations.
2. Norm bounds apply when it is easier to change the direction than the magnitude of an output.

Interpretation

PRINCETON UNIVERSITY

1. If $\mathcal{H}$ is very expressive, it is best to collapse all data points to the vertices of a standard simplex. This argument is not related to optimization algorithms, but the behavior is expected in the long term limit, as we approach minimal configurations.
2. Norm bounds apply when it is easier to change the direction than the magnitude of an output.
3. Euclidean geometry seems to play a role:
   - Radial Gaussian initialization
   - SGD Optimization

Interpretation

1. If $\mathcal{H}$ is very expressive, it is best to collapse all data points to the vertices of a standard simplex. This argument is not related to optimization algorithms, but the behavior is expected in the long term limit, as we approach minimal configurations.
2. Norm bounds apply when it is easier to change the direction than the magnitude of an output.
3. Euclidean geometry seems to play a role:
   - Radial Gaussian initialization
   - SGD Optimization
4. Generalization: Is $\mathcal{H}$ very expressive or just expressive enough for the problem?

Counterexamples for shallow network classifiers

Setting

We consider *binary classification*, i.e.

- $\xi : \mathbb{R}^d \to \{-1, 1\}$,
- $h : \mathbb{R}^d \to \mathbb{R}$, and

$$\begin{aligned}
\mathcal{R}(h) &= -\int_{\mathbb{R}^d} \log\left(\frac{\exp\left(\xi_x \cdot h(x)\right)}{\exp\left(h(x)\right) + \exp\left(-h(x)\right)}\right) \mathbb{P}(\mathrm{d}x) \\
&= \int_{\mathbb{R}^d} \log\left(1 + \exp\left(-2\,\xi_x \cdot h(x)\right)\right) \mathbb{P}(\mathrm{d}x) \\
&\approx \int_{\mathbb{R}^d} \exp\left(-2\,\xi_x \cdot h(x)\right) \mathbb{P}(\mathrm{d}x).
\end{aligned}$$

Geometry 1: ReLU activation

PRINCETON UNIVERSITY

We can take the infinite width limit of neural networks by replacing

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} a_i \, \sigma(w_i^T x + b_i)$$

with

$$f_\pi(x) = \mathbb{E}_{(a,w,b)\sim\pi}\left[a\,\sigma(w^T x + b)\right].$$

Geometry 1: ReLU activation

We can take the infinite width limit of neural networks by replacing

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} a_i \, \sigma(w_i^T x + b_i)$$

with

$$f_\pi(x) = \mathbb{E}_{(a,w,b)\sim\pi} \left[a \, \sigma(w^T x + b) \right].$$

Training weights (a_i, w_i, b_i) by gradient flow corresponds to training distribution π by Wasserstein gradient flow.

Geometry 1: ReLU activation

We can take the infinite width limit of neural networks by replacing

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} a_i\,\sigma(w_i^T x + b_i)$$

with

$$f_\pi(x) = \mathbb{E}_{(a,w,b)\sim\pi}\big[a\,\sigma(w^T x + b)\big].$$

Training weights (a_i, w_i, b_i) by gradient flow corresponds to training distribution π by Wasserstein gradient flow. Banach space with the norm

$$\|f\|_{\mathcal{B}} = \inf\left\{\mathbb{E}_\pi\big[|a|\,\big(|w| + |b|\big)\big] : \pi \text{ s.t. } f = f_\pi\right\}.$$

Barron space: Bach '16, E-Ma-Wu '17, E-Wojtowytsch '20, Siegel-Xu '21

Geometry 1: ReLU activation

Theorem (Chizat-Bach '20)

If π_0 is a sufficiently 'spread out' distribution and π_t is trained by (Wasserstein) gradient flow, then the following hold (under further conditions):

1. $\xi_x\, h_{\pi^t}(x) \to +\infty$ *for* $\mathbb{P}$*-almost every* x.
2. *There exist* $h^* \in \mathcal{B}$ *and* $\mu : [0,\infty) \to (0,\infty)$ *such that* $\mu(t)\, h_{\pi_t} \to h^*$ *locally uniformly on* $\mathbb{R}^d$.
3. *Let* $F : \mathcal{B} \to \mathbb{R}$, $F(h) = \min_{x \in \operatorname{spt} \mathbb{P}} \big(\xi_x \cdot h(x)\big)$. *Then* $h^* \in \operatorname{argmax}_{\|h\|_{\mathcal{B}} \leq 1} F$.

See also: Chizat-Bach '18, Wojtowytsch '20.

Geometry 1: ReLU activation

Consider a classification on the real line where

1. $C_{-1} \subseteq (-\infty, -1]$ and $C_1 \subseteq [1, \infty)$.
2. $-1 \in C_{-1}$ and $1 \in C_1$.

Lemma (E-W '20)

There exists a continuum of maximum margin classifiers

$$f_b(x) = \frac{1}{2+2b} \begin{cases} x+b & x > b \\ 2x & -b < x < b \\ x-b & x < -b \end{cases}, \qquad b \in [0,1].$$

Corollary (E-W '20)

If $C_{\pm 1}$ contains more than one point, f_b is not constant on the class.

Geometry 2: Non-convex data classes

PRINCETON UNIVERSITY

Consider

- $\mathbb{P} = p_1\,\delta_{-1} + p_2\,\delta_0 + p_3\,\delta_1$.
- $C_1 = \{-1, 1\}$ and $C_{-1} = \{0\}$.

Geometry 2: Non-convex data classes

PRINCETON UNIVERSITY

Consider

- $\mathbb{P} = p_1\,\delta_{-1} + p_2\,\delta_0 + p_3\,\delta_1$.
- $C_1 = \{-1, 1\}$ and $C_{-1} = \{0\}$.
- σ s.t. $\sigma(z) = 0$ for $z \leq 0$ and $\sigma(z) = 1$ for $z \geq 1$.

Geometry 2: Non-convex data classes

PRINCETON UNIVERSITY

Consider

- $\mathbb{P} = p_1\,\delta_{-1} + p_2\,\delta_0 + p_3\,\delta_1$.
- $C_1 = \{-1, 1\}$ and $C_{-1} = \{0\}$.
- σ s.t. $\sigma(z) = 0$ for $z \leq 0$ and $\sigma(z) = 1$ for $z \geq 1$.

Assume that at initialization

$$h(x) = a_1\,\sigma(-x) - a_2\sigma(x+1) + a_3\,\sigma(x)$$

for a_1, a_2, a_3.

Geometry 2: Non-convex data classes

Consider

- $\mathbb{P} = p_1\,\delta_{-1} + p_2\,\delta_0 + p_3\,\delta_1$.
- $C_1 = \{-1, 1\}$ and $C_{-1} = \{0\}$.
- σ s.t. $\sigma(z) = 0$ for $z \leq 0$ and $\sigma(z) = 1$ for $z \geq 1$.

Assume that at initialization

$$h(x) = a_1\,\sigma(-x) - a_2\sigma(x+1) + a_3\,\sigma(x)$$

for a_1, a_2, a_3. Since $\sigma' = 0$ $\mathbb{P}$-almost everywhere, the inner layer words do not evolve.

$$\begin{aligned}\mathcal{R}(a_1, a_2, a_3) &= \int_{\mathbb{R}} \exp\big(-\xi_x h(x)\big)\,\mathbb{P}(\mathrm{d}x)\\ &= p_1\,\exp(-a_1) + p_2\,\exp(-a_2) + p_3\,\exp(a_2 - a_3)\end{aligned}$$

Geometry 2: Non-convex data classes

The gradient flow equation

$$\begin{pmatrix} \dot{a}_1 \\ \dot{a}_2 \\ \dot{a}_3 \end{pmatrix} = \begin{pmatrix} p_1 \, \exp(-a_1) \\ p_2 \, \exp(-a_2) - p_3 \, \exp(a_2 - a_3) \\ p_3 \, \exp(a_2 - a_3) \end{pmatrix}$$

can be solved

Geometry 2: Non-convex data classes

The gradient flow equation

$$\begin{pmatrix} \dot{a}_1 \\ \dot{a}_2 \\ \dot{a}_3 \end{pmatrix} = \begin{pmatrix} p_1 \exp(-a_1) \\ p_2 \exp(-a_2) - p_3 \exp(a_2 - a_3) \\ p_3 \exp(a_2 - a_3) \end{pmatrix}$$

can be solved and

$$\lim_{t \to \infty} \left[f_{(a_1,a_2,a_3)(t)}(1) - f_{(a_1,a_2,a_3)(t)}(-1) \right] = \log \left(\frac{p_3}{2p_1} \right)$$

independently of a_1, a_2, a_3 at time $t = 0$.

Interpretation

PRINCETON UNIVERSITY

In shallow networks, neural collapse may not happen dynamically

- even in the output layer and
- even if the class is expressive enough to allow it.

Interpretation

In shallow networks, neural collapse may not happen dynamically

- even in the output layer and
- even if the class is expressive enough to allow it.

Two geometries:

- C_i intersects the convex hull of C_j or
- C_i and C_j are linearly separable and the activation is ReLU.

Interpretation

In shallow networks, neural collapse may not happen dynamically

- even in the output layer and
- even if the class is expressive enough to allow it.

Two geometries:

- C_i intersects the convex hull of C_j or
- C_i and C_j are linearly separable and the activation is ReLU.

Impact for deep learning:

- if classes are not 'geometrically nice' two layers before the output, they do not collapse in the output...
- ... especially when using a pre-trained model and adding few layers at the output.

Interpretation

In shallow networks, neural collapse may not happen dynamically

- even in the output layer and
- even if the class is expressive enough to allow it.

Two geometries:

- C_i intersects the convex hull of C_j or
- C_i and C_j are linearly separable and the activation is ReLU.

Impact for deep learning:

- if classes are not 'geometrically nice' two layers before the output, they do not collapse in the output...
- ... especially when using a pre-trained model and adding few layers at the output.

Related results: Mixon-Parshall-Pi '20, Lu-Steinerberger '21

Interpretation

PRINCETON UNIVERSITY

In shallow networks, neural collapse may not happen dynamically

- even in the output layer and
- even if the class is expressive enough to allow it.

Two geometries:

- C_i intersects the convex hull of C_j or
- C_i and C_j are linearly separable and the activation is ReLU.

Impact for deep learning:

- if classes are not 'geometrically nice' two layers before the output, they do not collapse in the output...
- ... especially when using a pre-trained model and adding few layers at the output.

Related results: Mixon-Parshall-Pi '20, Lu-Steinerberger '21

How does data become 'more separable' as it propagates through the layers of a DNN?

PRINCETON UNIVERSITY

Thank you for your attention.