Explicit construction of global minimizers and the interpretability problem in Deep Learning

Thomas Chen

University of Texas at Austin

Includes joint work with

Patricia Muñoz Ewald

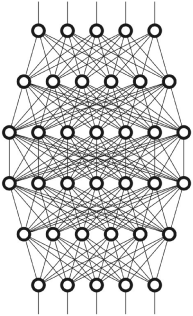
University of Texas at Austin

Texas A&M University, 2025

Financial support by NSF.

Deep Learning Networks

DL network for supervised learning: Inspired by brain architecture.



Input layer

L hidden layers

Output layer

Hidden layer $\sim$ affine map composed with a nonlinear activation fct.

Cost (loss) function on output layer, minimize over affine parameters.

Definition of DL network

Input layer with N_j training inputs for equivalence classes $j = 1, \dots, Q$.

$$x_{j,i}^{(0)} \in \mathbb{R}^M \quad , \quad i = 1, \dots, N_j$$

Hidden layers $\ell = 1, \dots, L$ with activation function σ (nonlinear !)

$$x_{j,i}^{(\ell)} = \sigma(W_\ell x_{j,i}^{(\ell-1)} + b_\ell) \quad \in \mathbb{R}^{M_\ell}$$

and affine map with (unknown) weight matrices and bias vectors

$$W_\ell \in \mathbb{R}^{M_\ell \times M_{\ell-1}} \quad , \quad b_\ell \in \mathbb{R}^{M_\ell}$$

Output layer

$$x_{j,i}^{(L+1)} = W_{L+1} x_{j,i}^{(L)} + b_{L+1} \quad \in \mathbb{R}^Q$$

Reference output vectors labeling j-th equivalence class

$$y_j \in \mathbb{R}^Q \quad , \quad j = 1, \dots, Q$$

Weighted $\mathcal{L}^2$ cost function with $\underline{N} := (N_1, \dots, N_Q)$

$$\mathcal{C}_{\underline{N}}[(W_i, b_i)_{i=1}^{L+1}] = \sum_{j=1,\dots,Q} \frac{1}{N_j} \sum_{i=1,\dots,N_j} \left| x_{j,i}^{(L+1)} - y_j \right|_{\mathbb{R}^Q}^2 .$$

Def: ReLU (Rectified Linear Unit) activation function σ.

Ramp function, acting component-wise

$$\sigma : A = [a_{ij}] \mapsto [(a_{ij})_+] \quad , \quad (a)_+ := \max\{0, a\}$$

Note that $\sigma(x) = x$ for $x \in \mathbb{R}^n_+$ and $\sigma(x) = 0$ for $x \in \mathbb{R}^n_-$.

Goal: Find cost minimizing weights, biases, to train DL network.

Zero loss minimizers W_i^*, b_i^* yield $\mathcal{C}_{\underline{N}}[(W_i^*, b_i^*)_{i=1}^{L+1}] = 0$.

Given new input, identifies its equivalence class.

Also often used: Entropy cost.

Gradient descent

Let $\underline{\theta} \in \mathbb{R}^K$ enlist components of all weights W_ℓ and biases b_ℓ:

$$K = \sum_{\ell=1}^{L+1} (M_\ell M_{\ell-1} + M_\ell) \ , \ \ M_0 \equiv M$$

Merge all vectors in output layer into

$$x_r[\underline{\theta}] := x_{j_r,i_r}^{(L+1)} \in \mathbb{R}^Q \ , \ \ \underline{x}[\underline{\theta}] := (x_1^T[\underline{\theta}], \dots, x_N^T[\underline{\theta}])^T \in \mathbb{R}^{QN}$$

Gradient descent method: Gradient flow of weights and biases

$$\partial_s \underline{\theta}(s) = -\nabla_{\underline{\theta}} \mathcal{C}[\underline{x}[\underline{\theta}(s)]] \ , \ \underline{\theta}(0) = \underline{\theta}_0 \ \in \mathbb{R}^K .$$

Monotone decreasing

$$\partial_s \mathcal{C}[\underline{x}[\underline{\theta}(s)]] = -\left| \nabla_{\underline{\theta}} \mathcal{C}[\underline{x}[\underline{\theta}(s)]] \right|_{\mathbb{R}^K}^2 \leq 0 \,,$$

$\mathcal{C}[\underline{x}[\underline{\theta}(s)]] \geq 0$ bounded below $\Rightarrow \mathcal{C}_* = \lim_{s\to\infty} \mathcal{C}[\underline{x}[\underline{\theta}(s)]]$ exists for any orbit $\{\underline{\theta}(s) | s \in \mathbb{R}\}$, and depends on the initial data $\underline{\theta}_0$.

Challenges of gradient descent method

Problems: Cost always converges to a stationary value, but not necessarily to global minimum. Typically, there may be many (approximate) local minima trapping the orbit ("landscape"), and identifying valid ones yielding sufficiently well-trained DL network relies on ad hoc methods getting flow unstuck from invalid ones.
In applications, $\underline{\theta}_0 \in \mathbb{R}^K$ often chosen at random.

Paradigm: Training data $(x_{j,i}^{(0)})_{j,i}$ generic $\Rightarrow \underline{x} : \mathbb{R}^K \to \mathbb{R}^{QN}$ generic.

- Underparametrized case: $K < QN$, embedding: Zero loss global minimum not reachable for generic training data distribution.
- Overparametrized case: $K \geq QN$, typically used. Can get zero loss global minimum if lucky.

[C-M Ewald '23] Non-generic $\Rightarrow$ zero loss in underparametrized DL exists.

Some related works

- T. Chen, J. Geom. Phys., 2004.
- W. E, Commun. Math. Stat., 2017.
- W. E, S. Wojtowytsch, Proc. MLR 2022.
- H. Gu, M. A. Katsoulakis, L. Rey-Bellet, B.J. Zhang, arXiv 2024.
- J.E. Grigsby, K. Lindsey, R. Meyerhoff, C. Wu, arXiv 2022.
- A. Jacot, F. Gabriel, C. Hongler, Adv. Neur. Inf. Proc. Sys. 2018.
- J.R. Lucas, J. Bae, M.R. Zhang, S. Fort, R. Zemel, R.B. Grosse, Proc. MLR 2021.

Neural collapse:
- V. Papyan, X.Y. Han, D. L. Donoho, Proc. NAS 2022.

Neural collapse

Prevalence of Neural Collapse during the terminal phase of deep learning training

Vardan Papyan[a,1], X.Y. Han[b,1], and David L. Donoho[a,2]

[a]Department of Statistics, Stanford University; [b]School of Operations Research and Information Engineering, Cornell University

This manuscript was compiled on August 24, 2020

Modern practice for training classification deepnets involves a *Terminal Phase of Training* (TPT), which begins at the epoch where training error first vanishes; During TPT, the training error stays effectively zero while training loss is pushed towards zero. Direct measurements of TPT, for three prototypical deepnet architectures and across seven canonical classification datasets, expose a pervasive inductive bias we call *Neural Collapse*, involving four deeply interconnected phenomena: (NC1) Cross-example within-class variability of last-layer training activations collapses to zero, as the individual activations themselves collapse to their class-means; (NC2) The class-means collapse to the vertices of a Simplex Equiangular Tight Frame (ETF); (NC3) Up to rescaling, the last-layer classifiers collapse to the class-means, or in other words to the Simplex ETF, i.e. to a *self-dual* configuration; (NC4) For a given activation, the classifier's decision collapses to simply choosing whichever class has the closest train class-mean, i.e. the Nearest Class-Center (NCC) decision rule. The symmetric and very simple geometry induced by the TPT confers important benefits, including better generalization performance, better robustness, and better interpretability.

Machine learning | Deep learning | Adversarial robustness | Simplex Equiangular Tight Frame | Nearest Class Center | Inductive bias

1. Introduction

Over the last decade, deep learning systems have steadily advanced the state-of-the-art in benchmark competitions, culminating in super-human performance in tasks ranging from image classification to language translation to game play. One might expect the trained networks to exhibit many particularities–making it impossible to find any empirical regularities across a wide range of datasets and architectures. On the contrary, in this article we present extensive measurements across image-classification datasets and architectures, exposing a common empirical pattern.

Our observations focus on today's standard training paradigm in deep learning, an accretion of several fundamental ingredients that developed over time: Networks are trained beyond zero misclassification error, approaching negligible cross-entropy loss, *interpolating* the in-sample training data; networks are *overparameterized*, making such memorization possible; and these parameters are layered in ever-growing *depth*, allowing for sophisticated feature engineering. A series of recent works (1–5) highlighted the paradigmatic nature of the practice of training well beyond zero-*error*, seeking zero-*loss*. We call the post-zero-error phase the **Terminal Phase of Training (TPT)**.

A scientist with standard preparation in mathematical statistics might anticipate that the linear classifier resulting from this paradigm, being a by-product of such training, would be quite arbitrary and vary wildly–from instance to instance, dataset to dataset, and architecture to architecture–thereby displaying no underlying cross-situational invariant structure. The scientist might further expect that the configuration of the fully-trained decision boundaries – and the underlying linear classifier defining those boundaries – would be quite arbitrary and vary chaotically from situation to situation. Such expectations might be supported by appealing to the overparameterized nature of the model, and to standard arguments whereby any noise in the data propagates during overparameterized training to generate disproportionate changes in the parameters being fit.

Defeating such expectations, we show here that TPT frequently induces an underlying mathematical simplicity to the trained deepnet model – and specifically to the classifier and last-layer activations – across many situations now considered canonical in deep learning. Moreover, the identified structure naturally suggests performance benefits. And indeed, we show that convergence to this rigid structure tends to occur simultaneously with improvements in the network's generalization performance as well as adversarial robustness.

We call this process **Neural Collapse**, and characterize it by four manifestations in the classifier and last-layer activations:

(NC1) Variability collapse: As training progresses, the within-class variation of the activations becomes negligible as these activations collapse to their class-means.

(NC2) Convergence to Simplex ETF: The vectors of the class-means (after centering by their global-mean)

Significance Statement

Modern deep neural networks for image classification have achieved super-human performance. Yet, the complex details of trained networks have forced most practitioners and researchers to regard them as blackboxes with little that could be understood. This paper considers in detail a now-standard training methodology: driving the cross-entropy loss to zero, continuing long after the classification error is already zero. Applying this methodology to an authoritative collection of standard deepnets and datasets, we observe the emergence of a simple and highly symmetric geometry of the deepnet features and of the deepnet classifier; and we document important benefits that the geometry conveys – thereby helping us understand an important component of the modern deep learning training paradigm.

V.P. and X.H. performed the experiments, formulated theorems, and derived proofs. D.D. guided the experimental design, formulated theorems, and derived proofs.

[1] V.P. and X.H. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: donoho@stanford.edu

www.pnas.org/cgi/doi/10.1073/pnas.XXXXXXXXXX

PNAS | August 24, 2020 | vol. XXX | no. XX | 1–17

Neural collapse

converge to having equal length, forming equal-sized angles between any given pair, and being the maximally pairwise-distanced configuration constrained to the previous two properties. This configuration is identical to a previously studied configuration in the mathematical sciences known as Simplex **Equiangular Tight Frame (ETF)** (6). See Definition 1.

(NC3) Convergence to self-duality: The class-means and linear classifiers – although mathematically quite different objects, living in dual vector spaces – converge to each other, up to rescaling. Combined with (NC2), this implies a *complete symmetry* in the network classifiers' decisions: each iso-classifier-decision region is isometric to any other such region by rigid Euclidean motion; moreover the class-means are each centrally located within their own specific regions, so there is no tendency towards higher confusion between any two classes than any other two.

(NC4) Simplification to Nearest Class-Center (NCC): For a given deepnet activation, the network classifier converges to choosing whichever class has the nearest train class-mean (in standard Euclidean distance).

We give a visualization of the phenomena (NC1)-(NC3) in Figure 1*, and define Simplex ETFs (NC2) more formally as follows:

Definition 1 (**Simplex ETF**). A *standard* Simplex ETF is a collection of points in $\mathbb{R}^C$ specified by the columns of

$$M^\star = \sqrt{\frac{C}{C-1}}\left(I - \frac{1}{C}\mathbb{1}\mathbb{1}^\top\right), \qquad [1]$$

where $I \in \mathbb{R}^{C\times C}$ is the identity matrix, and $\mathbb{1}_C \in \mathbb{R}^C$ is the ones vector. In this paper, we allow other poses, as well as rescaling, so the *general* Simplex ETF consists of the points specified by the columns of $M = \alpha U M^\star \in \mathbb{R}^{p\times C}$, where $\alpha \in \mathbb{R}_+$ is a scale factor, and $U \in \mathbb{R}^{p\times C}$ $(p \geq C)$ is a partial orthogonal matrix ($U^\top U = I$).

Properties (NC1)-(NC4) show that a highly symmetric and rigid mathematical structure with clear interpretability arises spontaneously during deep learning feature engineering, identically across many different datasets and model architectures.

(NC2) implies that the different feature means are 'equally spaced' around the sphere in their constructed feature space; (NC3) says the same for the linear classifiers in their own dual space; and moreover, that the linear classifiers are 'the same as' the class means, up to possible rescaling. These mathematical symmetries and rigidities vastly simplify the behavior and analysis of trained classifiers, as we show in Section 5 below, which contrasts the kind of qualitative understanding previously available from theory, against the precise and highly constrained predictions possible with (NC4).

(NC1)-(NC4) offer theoretically-established performance benefits: stability against random noise and against adversarial noise. And indeed, this theory bears fruit. We show that

*Figure 1 is, in fact, generated using *real measurements*, collected while training the VGG13 deepnet on CIFAR10: For three randomly selected classes, we extract the linear classifiers, class-means, and a subsample of twenty last-layer features at epochs 2, 16, 65, and 350. These entities are then rotated, rescaled, and represented in three-dimensions by leveraging the singular-value decomposition of the class-means. We omit further details as Figure 1 serves only to illustrate Neural Collapse on an abstract level.

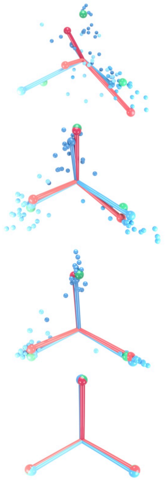


Fig. 1. Visualization of Neural Collapse: The figures depict, in three dimensions, Neural Collapse as training proceeds, from top to bottom. Green spheres represent the vertices of the standard Simplex ETF (Definition 1), red ball-and-sticks represent linear classifiers, blue ball-and-sticks represent class-means, and small blue spheres represent last-layer features. For all objects, we distinguish different classes via the shade of the color. As training proceeds, last-layer features collapse onto their class-means (NC1), class-means converge to the vertices of the Simplex ETF (NC2), the linear classifiers approach their corresponding class-means (NC3). An animation can be found here.

Explicit global cost minimization

[C-Munoz Ewald '23] Cost with j-th cluster average in output layer

$$\overline{x_j^{(L+1)}} = \frac{1}{N_j} \sum_{i=1}^{N_j} x_{j,i}^{(L+1)}$$

Result with explicit construction: Global minimization splits into

$$\begin{aligned} \mathcal{C}_{\underline{N}}[(W_i, b_i)_{i=1}^{L+1}] &= \sum_{j=1}^{Q} \frac{1}{N_j} \sum_{i=1}^{N_j} \left| x_{j,i}^{(L+1)} - y_j \right|_{\mathbb{R}^Q}^2 \\ &= \sum_{j=1}^{Q} \frac{1}{N_j} \sum_{i=1}^{N_j} \left| x_{j,i}^{(L+1)} - \overline{x_j^{(L+1)}} \right|_{\mathbb{R}^Q}^2 + \sum_{j=1}^{Q} \left| \overline{x_j^{(L+1)}} - y_j \right|_{\mathbb{R}^Q}^2 \\ &= \sum_{j=1}^{Q} \left(\frac{1}{N_j} \sum_{i=1}^{N_j} \left| \Delta x_{j,i}^{(L+1)} \right|_{\mathbb{R}^Q}^2 \right) + \sum_{j=1}^{Q} \left| \overline{x_j^{(L+1)}} - y_j \right|_{\mathbb{R}^Q}^2 \end{aligned}$$

Each of $L \geq Q$ hidden layers eliminates variance of one of Q clusters.

Output layer matches Q cluster averages to Q reference outputs y_j.

Truncation maps

Assuming all $W_\ell \in GL(Q)$ invertible, define *cumulative parameters*

$$\begin{aligned} W^{(\ell)} &:= W_\ell W_{\ell-1} \cdots W_1 \\ b^{(\ell)} &:= W_\ell \cdots W_2 b_1 + \cdots + W_2 b_{\ell-1} + b_\ell \\ \beta^{(\ell)} &:= (W^{(\ell)})^{-1} b^{(\ell)} \end{aligned} \tag{1}$$

for $\ell = 1, \dots, L$. Define affine maps and *truncation maps*

$$a^{(\ell)}(x) := W^{(\ell)} x + b^{(\ell)}$$

$$\begin{aligned} \tau^{(\ell)}(x) &:= (a^{(\ell)})^{-1} \circ \sigma \circ a^{(\ell)}(x) \\ &= (W^{(\ell)})^{-1} \sigma(W^{(\ell)}(x + \beta^{(\ell)})) - \beta^{(\ell)} . \end{aligned} \tag{2}$$

Composition property:

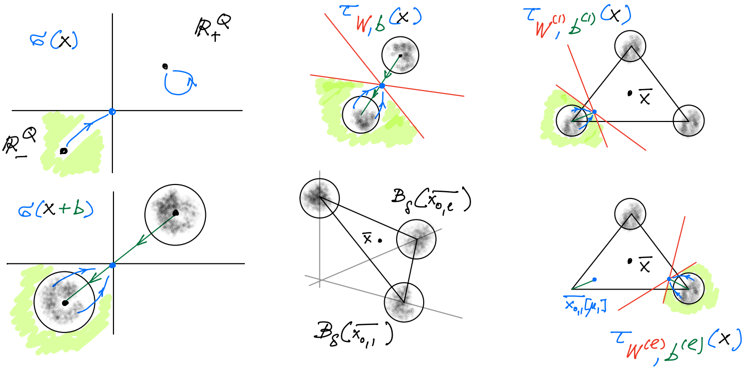
$$x^{(\ell)} = W^{(\ell)} \big(\tau^{(\ell)} \circ \cdots \circ \tau^{(1)}(x^{(0)}) + \beta^{(\ell)} \big)$$

The ℓ-th truncation maps is the pullback of the activation map in ℓ-th layer under $a^{(\ell)}$, and acts on the training data in the input layer.

Explicit global zero loss minimizers in underparametrized DL

Theorem [C-Munoz Ewald] $\exists$ explicit zero loss minimizers:

- Recursively reduce j-th cluster of training data to point $\overline{x_{0,j}}[\mu_j]$ where $\mu_j \in \mathbb{R}$ parametrizes distance from cluster center to barycenter $\overline{x}$.
- Obtain Q distinct points $\{\overline{x_{0,j}}[\mu_j]\}_{j=1}^{Q}$ in output layer.
- Minimize cost explicitly by matching them to $y_1, \ldots, y_Q$.



Cost is bounded by deviations in barycentric coordinates

Theorem (C-Muñoz Ewald 2023)

The cost satisfies the upper bound (least square in W_{L+1}, b_{L+1})

$$\min_{\underline{W}^{(L)}, W_{L+1}, \underline{b}^{(L)}, b_{L+1}} \mathcal{C}_{\underline{N}}[\underline{W}^{(L)}, W_{L+1}, \underline{b}^{(L)}, b_{L+1}] \leq C \min_{\underline{W}^{(L)}, \underline{b}^{(L)}} \delta_P \,,$$

with deviation w.r.t. truncated cluster centers in barycentric coordinates ,

$$\delta_P := \sup_{j,i} \left| [\overline{x_{0,1}}[\mu_1] \cdots \overline{x_{0,Q}}[\mu_Q]]^{-1} \Delta\tau^{(L)}(x_{j,i}^{(0)}) \right|$$

$$\overline{x_{0,j}}[\mu_j] := \frac{1}{N_j} \sum_{i=1}^{N_j} \tau^{(L)}(x_{j,i}^{(0)}) \quad , \quad \Delta\tau^{(L)}(x_{j,i}^{(0)}) := \tau^{(L)}(x_{j,i}^{(0)}) - \overline{x_{0,j}}[\mu_j]$$

δ_P measures the signal to noise ratio of the truncated training input data.
Invariant under $GL(Q)$ action in input space (incl. scalings and rotations).

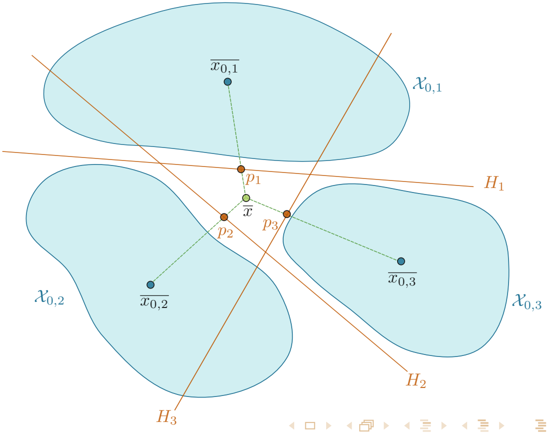
Theorem (C-Muñoz Ewald 2024)

Arbitrary non-increasing layer dimensions, data sequentially linearly separable by hyperplanes. For Q classes of data in R^M, $L \geq Q$ hidden layers, global zero loss minimizers with $Q(M+2)$ parameters.

Unique ordering

$1 \to 2 \to 3$

Sequential application of conical approximation to support vector machine



Derivation of effective gradient flow equations

[C '25] Standard $\mathcal{L}^2$ cost with $\underline{\tau}^{(L)} := \tau^{(L)} \circ \cdots \circ \tau^{(1)}$

$$\mathcal{C}_{\underline{N}} = \frac{1}{2} \sum_{j=1}^{Q} \frac{1}{N_j} \sum_{i=1}^{N_j} \left| W^{(L+1)} \big(\underline{\tau}^{(L)}(x_{j,i}^{(0)}) - (W^{(L+1)})^{-1} y_j \big) \right|_{\mathbb{R}^Q}^2 \tag{3}$$

Pullback metric in input space via map $W^{(L+1)}$: input $\to$ output space.

Non-Euclidean, time dependent metric introduces many of the known complications ("cost landscape").

Here, we propose to investigate the Euclidean $\mathcal{L}^2$ cost in the input space,

$$\widetilde{\mathcal{C}_{\underline{N}}} := \frac{1}{2} \sum_{j=1}^{Q} \frac{1}{N_j} \sum_{i=1}^{N_j} \left| \underline{\tau}^{(L)}(x_{j,i}^{(0)}) - (W^{(L+1)})^{-1} y_j \right|_{\mathbb{R}^Q}^2 \tag{4}$$

Study the gradient flow at fixed $W^{(L+1)}$ (quite common).

Derivation of effective gradient flow equations

Observe: Activation σ distinguishes specific coordinate system !
Polar decomposition of the cumulative weight

$$W^{(\ell)} = |W^{(\ell)}| R_\ell$$

with $R_\ell \in O(Q)$ orthogonal, and $|W^{(\ell)}|$ symmetric. Accordingly,

$$|W^{(\ell)}| = \widetilde{R}_\ell^T W_*^{(\ell)} \widetilde{R}_\ell \quad \text{with } W_*^{(\ell)} \geq 0 \text{ diagonal}$$

$\widetilde{R}_\ell \in SO(Q)$ freedom of rotating coordinate system in which σ is defined.

Choose the cumulative weights *adapted to activation* in that $\widetilde{R}_\ell = \mathbf{1}$,

$$W^{(\ell)} = W_*^{(\ell)} R_\ell$$

Then, truncation maps $\tau^{(\ell)}$ independent of $W_*^{(\ell)}$, due to

$$(W_*^{(\ell)})^{-1} \sigma(W_*^{(\ell)} x) = \sigma(x)$$

Thus, $\beta^{(\ell)} \in \mathbb{R}^Q$ and $R_\ell \in O(Q)$ parametrize the DL network.

Derivation of effective gradient flow equations

Empirical probability distribution for ℓ-th cluster of training inputs

$$\mu_\ell(x) := \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \delta(x - x_{\ell,i}^{(0)}) \,,$$

where δ is the Dirac delta distribution. Let

$$\widetilde{y}_\ell := (W^{(Q+1)})^{-1} y_\ell$$

for brevity, with $W^{(Q+1)}$ fixed.

Cluster separated truncations: $\tau^{(\ell)}$ acts nontrivially only on training inputs in the ℓ-th cluster.
On all other clusters, $\tau^{(\ell)}(x_{\ell',i}^{(0)}) = x_{\ell',i}^{(0)}$ for all $\ell' \neq \ell$, acts as identity.
Crucial for explicit construction of global cost minimizers for underparametrized DL in [C-Muñoz Ewald 2023].

Derivation of effective gradient flow equations

Theorem [C '25] Effective equations for $\beta^{(\ell)}(s)$ and $R_\ell(s)$

$$\begin{aligned} \partial_s(\beta^{(\ell)} + \widetilde{y}_\ell) &= -R_\ell^T J_0^{(\ell)\perp} R_\ell(\beta^{(\ell)} + \widetilde{y}_\ell) \\ \partial_s R_\ell &= -\Omega_\ell R_\ell \end{aligned} \tag{5}$$

where

$$J_0^{(\ell)\perp} = \int_{\mathbb{R}^Q \setminus \mathbb{R}^Q_+} dx\, \mu_\ell(a^{-1}_{R_\ell,\beta^{(\ell)}}(x)) H^\perp(x)\,,$$

is a diagonal matrix with

$$H^\perp(x) = \mathbf{1}_{Q\times Q} - H(x) \quad , \quad H(x) = \mathrm{diag}(h(x_i))$$

and

$$\Omega_\ell = \int_{\mathbb{R}^Q \setminus (\mathbb{R}^Q_+ \cup \mathbb{R}^Q_-)} dx\, \mu_\ell(a^{-1}_{R_\ell,\beta^{(\ell)}}(x)) \big[H(x)\,,\; M^{(\ell)}(x) \big]\,,$$

where $[A, B] = AB - BA$ is the commutator of $A, B \in \mathbb{R}^{Q\times Q}$, and

$$M^{(\ell)}(x) := \frac{1}{2}\Big(x(\beta^{(\ell)} + \widetilde{y}_\ell)^T R_\ell^T + R_\ell(\beta^{(\ell)} + \widetilde{y}_\ell) x^T \Big)\,.$$

Explicit solutions

Proposition [C '25] Explicit solutions.

- The pair $(\beta^{(\ell)}, R_\ell)$ is an equilibrium solution if

$$\mathrm{supp}\Big(\mu_\ell \circ a^{-1}_{R_\ell,\beta^{(\ell)}}\Big) \subset \mathbb{R}^Q_+ \,, \tag{6}$$

and $\tau^{(\ell)}$ acts as the identity on the ℓ-th cluster, or

$$\mathrm{supp}\Big(\mu_\ell \circ a^{-1}_{R_\ell,\beta^{(\ell)}}\Big) \subset \mathbb{R}^Q_- \,, \tag{7}$$

and ℓ-th cluster is contracted to a point.

- If the initial data $(\beta^{(\ell)}(0), R_\ell(0))$ is such that

$$\mathrm{supp}\Big(\mu_\ell \circ a^{-1}_{R_\ell(0),\beta^{(\ell)}(0)}\Big) \cap \mathbb{R}^Q \setminus (\mathbb{R}^Q_+ \cup \mathbb{R}^Q_-) \neq \emptyset \,, \tag{8}$$

and the support of $\mu_\ell \circ a^{-1}_{R_\ell,\beta^{(\ell)}}$ is concentrated in $\mathbb{R}^Q_-$, s.t. for $\eta > 0$ small,

$$J_0^{(\ell)\perp} > 1 - \eta \tag{9}$$

then (up to technical assumptions) the following holds.

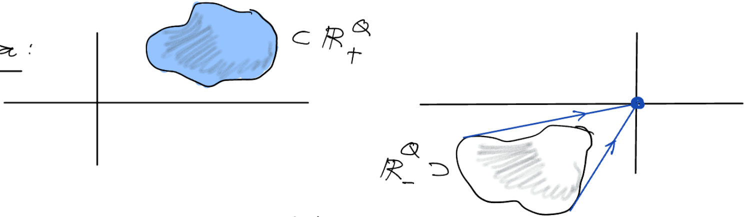
Derivation of effective gradient flow equations

- Solution of gradient flow translates $\mu_\ell \circ a^{-1}_{R_\ell(s),\beta^{(\ell)}(s)}$ into $\mathbb{R}^Q_-$ in finite time $s = s_1 < \infty$.
- $\beta^{(\ell)}(s) \to -\widetilde{y}_\ell$ exponentially as $s \to \infty$.
- For $s > s_1$, the weight matrix $R_\ell(s) = R_\ell(s_1)$ is stationary. In particular, this implies that the entire ℓ-th cluster is collapsed into the point $\beta^{(\ell)}(s)$ for $s > s_1$.

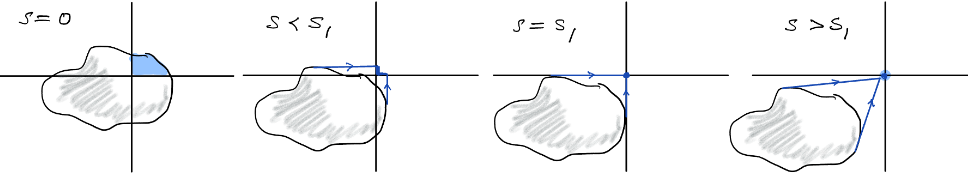
Provides dynamical interpretation of neural collapse on level of training data in input space, see [Papyan-Han-Donoho], [C-Ewald].

Solutions to effective gradient flow equations

Equilibria:



Partially truncated initial data:



Geometric structure of overparametrized DL networks

Vector $\underline{\theta} \in \mathbb{R}^K$ of components of all weights W_ℓ and biases b_ℓ,

$$K = \sum_{\ell=1}^{L+1} (M_\ell M_{\ell-1} + M_\ell)$$

In the output layer, we define

$$x_r[\underline{\theta}] := x_{j_r,i_r}^{(L+1)} \in \mathbb{R}^Q \quad , \quad \underline{x}[\underline{\theta}] := (x_1^T[\underline{\theta}], \dots, x_N^T[\underline{\theta}])^T \in \mathbb{R}^{QN}$$

Map $\omega : \{1, \dots, N\} \to \{1, \dots, Q\}$: Input $x_r^{(0)}$ assigned to output $y_{\omega(r)}$.

$$\underline{y}_\omega := (y_{\omega(1)}^T, \dots, y_{\omega(N)}^T)^T \in \mathbb{R}^{NQ}$$

Then, $\mathcal{L}^2$ cost is (assume all N_ℓ equal)

$$\mathcal{C}[\underline{x}[\underline{\theta}]] \quad = \quad \frac{1}{2N} \left| \underline{x}[\underline{\theta}] - \underline{y}_\omega \right|_{\mathbb{R}^{QN}}^2$$

Key observation: Cost depends on $\underline{\theta}$ only via $\underline{x}[\underline{\theta}]$.

Jacobian matrix for $f : \mathbb{R}^K \to \mathbb{R}^{QN}$, $\underline{\theta} \mapsto \underline{x}[\underline{\theta}]$

$$D[\underline{\theta}] \quad := \quad \left[\frac{\partial x_j[\underline{\theta}]}{\partial \theta_\ell}\right] = \begin{bmatrix} \frac{\partial x_1[\underline{\theta}]}{\partial \theta_1} & \cdots & \frac{\partial x_1[\underline{\theta}]}{\partial \theta_K} \\ \cdots & \cdots & \cdots \\ \frac{\partial x_N[\underline{\theta}]}{\partial \theta_1} & \cdots & \frac{\partial x_N[\underline{\theta}]}{\partial \theta_K} \end{bmatrix} \in \mathbb{R}^{QN \times K}$$

Therefore, Euclidean (!) gradient flow for $\underline{\theta}(s)$ can be written as

$$\partial_s \underline{\theta}(s) = -\nabla_{\underline{\theta}} \mathcal{C}[\underline{x}[\underline{\theta}]] = -D^T[\underline{\theta}(s)] \nabla_{\underline{x}} \mathcal{C}[\underline{x}[\underline{\theta}(s)]] \, .$$

Moreover, $\partial_s \underline{x}[\underline{\theta}(s)] = -D[\underline{\theta}(s)] \partial_s \underline{\theta}(s)$.

Induced gradient flow in output layer for $\underline{x}(s) := \underline{x}[\underline{\theta}(s)]$

$$\partial_s \underline{x}(s) = -(DD^T)[\underline{\theta}(s)] \, \nabla_{\underline{x}} \mathcal{C}[\underline{x}(s)] \quad \in \mathbb{R}^{QN}$$

Because $\mathrm{rank} DD^T \leq \min\{K, QN\}$
$\Rightarrow K \geq QN$ necessary for invertibility, overparametrized DL.
If invertible, $DD^T \nabla_{\underline{x}} =$ gradient w.r.t Riemannian metric $(DD^T)^{-1}$.
Metric $(DD^T)^{-1}$ on $\mathbb{R}^{QN}$ is source of complicated "energy landscape" !

Trapping of orbits

Assume $DD^T > 0$ full rank, but $DD^T > \lambda$ for $\lambda \ll 1$ or $\nexists$ such $\lambda > 0$.

There are no local equilibria

$$0 = \underbrace{(DD^T)[\underline{\theta}_*]}_{invertible} \nabla_{\underline{x}} \mathcal{C}[\underline{x}[\underline{\theta}_*]] \implies \underbrace{\nabla_{\underline{x}} \mathcal{C}[\underline{x}[\underline{\theta}_*]] = \frac{1}{N}(\underline{x}[\underline{\theta}_*] - \underline{y}_\omega) = 0}_{global\ minimum}$$

Proposition (C'24, trapping of orbits)

Assume $\exists U \subset \mathbb{R}^K$ *region and* $\epsilon > 0$ *such that for all* $\underline{\theta} \in U$

$$|\, D^T \nabla_{\underline{x}} \mathcal{C}[\underline{x}[\underline{\theta}]]\,|_{\mathbb{R}^K} \; < \; \epsilon\,|\, \nabla_{\underline{x}} \mathcal{C}[\underline{x}[\underline{\theta}]]\,|_{\mathbb{R}^{QN}}$$

Let $I = \{s \in \mathbb{R}_+ | \underline{\theta}(s) \in U\}$ *with* $s_0 = \inf I$ *and* $L_U := |\{\underline{\theta}(s) | s \in I\} \cap U|$.

$$\implies \quad |I| \; > \; \frac{N\,L_U}{|\underline{x}[\underline{\theta}(s_0)] - \underline{y}_\omega|} \, \frac{1}{\epsilon}$$

Proof. Arc length

$$
\begin{aligned}
L_U &= \Big| \{\underline{\theta}(s) | s \in \mathbb{R}_+\} \cap U \Big| \\
&= \int_I ds \, | \nabla_{\underline{\theta}} \mathcal{C}[\underline{x}[\underline{\theta}(s)]] \, | \\
&\leq |I| \, \epsilon \, \sup_{s \in I} | \nabla_{\underline{x}} \mathcal{C}[\underline{x}[\underline{\theta}(s)]] \, | \\
&= |I| \, \epsilon \left(\frac{2}{N} \sup_{s \in I} | \mathcal{C}[\underline{x}[\underline{\theta}(s)]] \, | \right)^{\frac{1}{2}} \\
&= |I| \, \epsilon \left(\frac{2}{N} \, | \mathcal{C}[\underline{x}[\underline{\theta}(s_0)]] \, | \right)^{\frac{1}{2}} \\
&= \frac{|I| \epsilon}{N} \, \Big| \underline{x}[\underline{\theta}(s_0)] - \underline{y}_{\omega} \Big|
\end{aligned}
$$

where $s_0 = \inf I$, using monotone decrease of cost along orbit. □

Differential geometry: Definition of gradient requires choice of metric.
Key insight: Instead of picking Euclidean metric in parameter space $\mathbb{R}^K$, choose Euclidean metric in output layer, and pull it back to $\mathbb{R}^K$.

Theorem (C 2024)

Assume the overparametrized case $K \geq QN$, and that

$$\mathrm{rank}(D[\underline{\theta}]) = QN$$

is maximal in the region $\underline{\theta} \in U \subset \mathbb{R}^K$. Let

$$\mathrm{Pen}[D] := D^T (DD^T)^{-1} \quad \in \mathbb{R}^{K \times QN}$$

Penrose inverse of $D[\underline{\theta}]$ for $\underline{\theta} \in U$, generalizes matrix inverse by way of

$$\mathrm{Pen}[D]D = P \quad , \quad D\mathrm{Pen}[D] = \mathbf{1}_{QN \times QN}$$

$P = P^2 = P^T \in \mathbb{R}^{K \times K}$ orthoprojector onto range of $D^T \in \mathbb{R}^{K \times QN}$.

Theorem (C 2024, continued)

If $\underline{\theta}(s) \in U$ is a solution of the modified gradient flow

$$\partial_s \underline{\theta}(s) = -\underbrace{\mathrm{Pen}[D[\underline{\theta}(s)]]\,\mathrm{Pen}[D^T[\underline{\theta}(s)]]}_{=(DD^T)^+ \text{ generalized inverse}} \nabla_{\underline{\theta}} \mathcal{C}[\underline{x}[\underline{\theta}(s)]]$$

then $\underline{x}(s) = \underline{x}[\underline{\theta}(s)] \in \mathbb{R}^{QN}$ is equivalent to Euclidean gradient flow

$$\partial_s \underline{x}(s) = -\nabla_{\underline{x}} \mathcal{C}[\underline{x}(s)] \ , \ \underline{x}(0) = \underline{x}[\underline{\theta}_0] \in \mathbb{R}^{QN} .$$

In particular, along any orbit $\underline{\theta}(s) \in U$, $s \in \mathbb{R}_+$,

$$\mathcal{C}[\underline{x}[\underline{\theta}(s)]] = e^{-\frac{2s}{N}} \mathcal{C}[\underline{x}[\underline{\theta}_0]] \ , \ \underline{x}[\underline{\theta}(s)] = \underline{y}_\omega + e^{-\frac{s}{N}} (\underline{x}(\underline{\theta}_0) - \underline{y}_\omega) ,$$

at uniform exponential convergence rates.

Pullback bundle with induced bundle metric on $\mathbb{R}^K$ and bundle gradient. Relationship to sub-Riemannian geometry.

Relation to sub-Riemannian geometry

Invariant geometric meaning: Assume $K > QN$ overparametrized

Then, with $f : \mathbb{R}^K \to \mathbb{R}^{QN}$, $\underline{\theta} \mapsto \underline{x}[\underline{\theta}]$,

$$\mathcal{V} := f^* T\mathbb{R}^{QN} \subset T\mathbb{R}^K$$

pullback vector bundle of fiber dimension QN.

Pullback bundle metric for sections $V, W \in \Gamma(T\mathbb{R}^K)$

$$h(V, W) = \langle f_* V, f_* W \rangle_{T\mathbb{R}^{QN}}$$

Bundle gradient of $F : \mathbb{R}^K \to \mathbb{R}$

$$dF(V) = h(V, \mathrm{grad}_h(F))$$

Then, with Jacobi matrix $D \equiv Df$, coordinate representation

$$\mathrm{grad}_h(F) = \mathrm{Pen}[D]\mathrm{Pen}[D^T]\nabla_{\underline{\theta}} F$$

In general, triple $(\mathbb{R}^K, \mathcal{V}, h)$ is a *sub-Riemannian manifold*.

Euclidean gradient flow in output layer with $s \in \mathbb{R}_+$,

$$\partial_s \underline{x}(s) = -\nabla_{\underline{x}} \mathcal{C}[\underline{x}(s)] \ , \ \underline{x}(0) \in \mathbb{R}^{QN} \ , \ \text{with } \mathcal{C}[\underline{x}] = \frac{1}{2N} |\underline{x} - \underline{y}_\omega|^2$$

Equivalent to

$$\begin{aligned} \partial_s(\underline{x}(s) - \underline{y}_\omega) &= -\frac{1}{N}(\underline{x}(s) - \underline{y}_\omega) \\ \Rightarrow \quad \underline{x}(s) - \underline{y}_\omega &= e^{-\frac{s}{N}}(\underline{x}(0) - \underline{y}_\omega) \\ \Rightarrow \quad \mathcal{C}[\underline{x}(s)] &= e^{-\frac{2s}{N}} \mathcal{C}[\underline{x}(0)] \, . \end{aligned}$$

Exponential convergence rates are uniform w.r.t. initial data.

$$\underline{x}_* := \lim_{s \to \infty} \underline{x}(s) = \underline{y}_\omega$$

unique global minimizer of the $\mathcal{L}^2$ cost, by convexity of $\mathcal{C}$ in $\underline{x} - \underline{y}_\omega$.

Theorem (C'24, overparametrized with rank loss)

Assume $\mathrm{rank}(D) \leq QN$. *Then, standard gradient flow yields*

$$\partial_s \underline{x}(s) \;=\; -(\mathcal{P}DD^T\mathcal{P})[\underline{\theta}(s)]\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}(s)]]$$

with $\underline{x}(0) = \underline{x}[\underline{\theta}_0]$, *and* $\mathcal{P}$ *orthoprojector onto* $\mathrm{range}(DD^T)$ *in* $\mathbb{R}^{QN}$.

Generalized adapted flow: Define differential-algebraic system

$$\begin{aligned}
\partial_s \underline{\theta}(s) &= D^T[\underline{\theta}(s)]\Psi[\underline{\theta}(s)] \\
\Psi[\underline{\theta}(s)] &= \mathrm{argmin}_{\Psi}\{\, |D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]\Psi + \nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}(s)]]\,|^2_{\mathbb{R}^{QN}} \,\} \\
\underline{\theta}(0) &= \underline{\theta}_0 \in \mathbb{R}^K .
\end{aligned}$$

That is, $\Psi[\underline{\theta}(s)]$ *solves via least square optimization*

$$D[\underline{\theta}(s)]D^T[\underline{\theta}(s)]\Psi = -\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}(s)]] \;+\; \text{minimal error in } L^2 .$$

Then, $\underline{x}(s) = \underline{x}[\underline{\theta}(s)]$ *with* $\underline{x}(0) = \underline{x}[\underline{\theta}_0]$ *solves*

$$\partial_s \underline{x}(s) \;=\; -\mathcal{P}[\underline{\theta}(s)]\nabla_{\underline{x}}\mathcal{C}[\underline{x}[\underline{\theta}(s)]] .$$

Theorem (C'24, overparametrized with rank loss)

$\underline{\theta}_* \in \mathbb{R}^K$ is equilibrium of the standard gradient flow
$\Longleftrightarrow$ $\underline{\theta}_*$ is equilibrium of the geometrically adapted gradient flow.

Assume activation function σ smooth. If $\mathrm{rank}(D) = r < QN$ in $U \subset \mathbb{R}^K$, then any local equilibrium $\underline{\theta}_*$ is contained in an $(K-r)$-dimensional critical submanifold $\mathcal{M}_{crit} \subset U$, generically in the sense of Sard.

Proof. Assume $V_\alpha : \mathbb{R}^K \to \mathbb{R}^{QN}$, $\alpha = 1, \dots, r$, are linearly independent column vectors of D. Obtain family of smooth functions

$$\begin{aligned} g_\alpha[\underline{\theta}] &:= \left\langle V_\alpha[\underline{\theta}] \, , \, \mathcal{P}[\underline{\theta}] \nabla_{\underline{x}} \mathcal{C}[\underline{x}[\underline{\theta}]] \right\rangle_{\mathbb{R}^{QN}} \\ &= \left\langle V_\alpha[\underline{\theta}] \, , \, \nabla_{\underline{x}} \mathcal{C}[\underline{x}[\underline{\theta}]] \right\rangle_{\mathbb{R}^{QN}} \, , \; \alpha = 1, \dots, r \end{aligned}$$

By Sard's theorem, set of equilibrium solutions in $U \subset \mathbb{R}^K$

$$\mathcal{M}_{crit} = U \cap \bigcap_{\alpha=1}^{r} g_\alpha^{-1}(0) \, .$$

is generically a $(K-r)$-dimensional submanifold of U. □

Theorem (C-Ewald 2024)

Standard and modified gradient flow have same critical points, and

$$\partial_s \underline{\theta}(s) = -\Big((1-\alpha) + \alpha \mathrm{Pen}[D[\underline{\theta}(s)]]\,\mathrm{Pen}[D^T[\underline{\theta}(s)]]\Big)\nabla_{\underline{\theta}}\mathcal{C}[\underline{x}[\underline{\theta}(s)]]\,,$$

establishes a homotopy equivalence of flows parametrized by $\alpha \in [0,1]$.

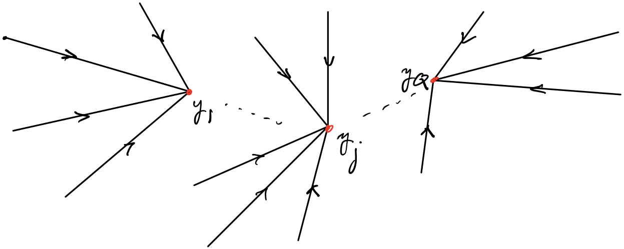
If D *has full rank, then the time reparametrization* $t = 1 - e^{-s/N}$,

$$\widetilde{\underline{x}}(t) := \underline{x}[\underline{\theta}(\underbrace{-N\ln(1-t)}_{=s(t)})]$$

maps the flow at $\alpha = 1$ *to linear interpolation in output space*

$$\widetilde{\underline{x}}(t) = (1-t)\underline{x}_0 + t\underline{y}_{\omega} \;\;,\;\; \widetilde{\underline{x}}(0) = \underline{x}[\underline{\theta}_0] \in \mathbb{R}^{QN}\,.$$

Standard gradient flow is homotopy and reparametrization equivalent to linear flow on straight lines towards reference outputs



Neural collapse: Cluster variances converge to zero, cluster averages converge to reference outputs, at uniform exponential rate in geometrically adapted flow.

Reparametrized Euclidean flow under rank loss

Proposition (C-Ewald 2024)

Assume rank loss, $\mathrm{rank}(DD^T) < QN$.
Then, reparametrized Euclidean gradient flow in output space satisfies

$$\partial_t \underline{\widetilde{x}}(t) = -\frac{1}{1-t}\mathcal{P}_t(\underline{\widetilde{x}}(t) - \underline{y}) \quad , \quad \underline{\tilde{x}}(0) = \underline{x}_0 \quad , \quad t \in [0,1)\,,$$

where $\underline{\widetilde{x}}(t) = \underline{x}[\underline{\theta}(s(t))]$ *and* $\mathcal{P}_t := \mathcal{P}_{\mathrm{range}(DD^T)}[\underline{\theta}(s(t))]$.

Deviation from linear interpolation

$$\underline{\widetilde{x}}(t) - ((1-t)\underline{x}_0 + t\underline{y}) = \int_0^t dt'\,\mathcal{U}_{t,t'}\,\frac{1-t}{1-t'}\,\mathcal{P}^{\perp}_{t'}\,(\underline{x}_0 - \underline{y})$$

with linear propagator $\mathcal{U}_{t,t'}$

$$\partial_t \mathcal{U}_{t,t'} = \frac{1}{1-t}\,\mathcal{P}_t\,\mathcal{U}_{t,t'} \quad , \quad \mathcal{U}_{t',t'} = \mathbf{1}_{Q\times Q} \quad , \quad t,t' \in [0,1)\,.$$

Conclusion

Underparametrized DL

Zero loss global cost minimizers exist for non-generic training data distributions, such as clustered data. Explicit construction through complexity reduction via truncation maps. Hidden layers eliminate cluster variances via truncation maps, and output layer matches cluster averages to reference outputs.

Overparametrized DL

Exploit arbitrariness of choice of Riemannian structure in definition of gradient. Construct geometrically adapted gradient flow inducing Euclidean gradient flow in output layer with uniform convergence rates. If Jacobian D has full rank, then standard gradient flow is homotopy and reparametrization equivalent to linear interpolation in output space.

Neural collapse occurs in both cases, but for different reasons !

Thank you for your attention !