Explicit construction of global minimizers in Deep Learning networks

Thomas Chen

University of Texas at Austin

Joint work with

Patricia Muñoz Ewald

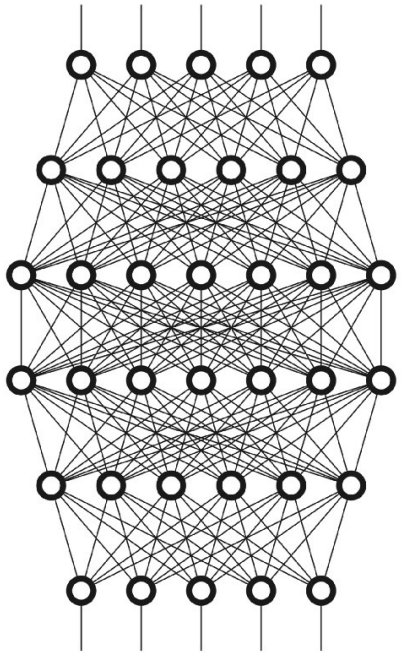
University of Texas at Austin

AMS Sectional Meeting, UT San Antonio 2024

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

Reference output vectors

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

Input layer with N_j training inputs belonging to y_j, for each j

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

Hidden layers $\ell = 1, \ldots, L$

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

parametrized by (unknown) weight matrices and bias vectors

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

ReLU activation function σ (nonlinear !), acting component-wise

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

Output layer

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

Weighted cost function

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

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

Gradient descent

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

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

Vector in output layer

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

Gradient descent method: Gradient flow of weights and biases

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

Monotone decreasing

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

$\mathcal{C}[\underline{x}[\underline{\theta}(s)]] \geq 0$ 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: The cost always converges to a stationary value, but not necessarily to the global minimum. Typically, there may be many (approximate) local minima trapping the orbit ("landscape"), and identifying valid ones yielding a 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.

- Underparametrized case: $K < QN$, gradient descent generically can't find global minimum.
- Overparametrized case: $K \geq QN$, typically used. Can get global minimum if lucky.

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:

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

Construction of global minimizers in underparametrized DL

Joint work with Patricia Muñoz Ewald, 2023.

Assume $M = M_\ell = Q$.

Define cluster average of all training inputs belonging to output y_j,

$$\overline{x_{0,j}} := \frac{1}{N_j} \sum_{i=1}^{N_j} x_{j,i}^{(0)} \in \mathbb{R}^Q \quad , \quad j = 1, \ldots, Q.$$

Define deviations from $\overline{x_{0,j}}$ belonging to output y_j

$$\Delta x_{j,i}^{(0)} := x_{j,i}^{(0)} - \overline{x_{0,j}} \, .$$

Assume $\overline{X_0^{red}} := [\overline{x_{0,1}} \cdots \overline{x_{0,Q}}] \in GL(Q)$ and data sufficiently clustered,

$$| \, (\overline{X_0^{red}})^{-1} \Delta x_{j,i}^{(0)} \, | < \frac{1}{4}$$

Deviation in barycentric coordinates associated to simplex spanned by cluster centers $\overline{x_{0,j}}$ (in barycentric coords, $\overline{x_{0,j}} \sim (0, ..., 1, ..., 0)$).

Main insight: Instead of focusing on $\sigma(Wx+b)$ in hidden layer

Definition (C-Muñoz Ewald 2023)

Given $W \in GL(Q)$, $b \in \mathbb{R}^Q$, define the *truncation map*

$$\begin{aligned} \tau_{W,b} \; : \; \mathbb{R}^Q \; &\to \; \mathbb{R}^Q \\ x \; &\mapsto \; W^{-1}(\sigma(Wx+b)-b)\,, \end{aligned}$$

$\tau_{W,b} = a_{W,b}^{-1} \circ \sigma \circ a_{W,b}$ under affine map $a_{W,b} : x \mapsto Wx+b$.

Proposition (C-Muñoz Ewald 2023)

Recursively, for $\ell = 1, \dots, L$,

$$\begin{aligned} x_{j,i}^{(\ell)} &= W_\ell \, \tau_{W_\ell, b_\ell}(x_{j,i}^{(\ell-1)}) + b_\ell \\ &= \cdots = W^{(\ell)} \, \tau_{\underline{W}^{(\ell)}, \underline{b}^{(\ell)}}(x_{j,i}^{(0)}) + b^{(\ell)} \end{aligned}$$

where (recursive structure similar to renormalization map in QFT)

$$\begin{aligned} &\tau_{\underline{W}^{(\ell)}, \underline{b}^{(\ell)}}(x_{j,i}^{(0)}) \\ &:= \tau_{W^{(\ell)}, b^{(\ell)}}(\tau_{W^{(\ell-1)}, b^{(\ell-1)}}(\cdots \tau_{W^{(2)}, b^{(2)}}(\tau_{W^{(1)}, b^{(1)}}(x_{j,i}^{(0)}))\cdots)) \end{aligned}$$

for cumulative weights and biases

$$\begin{aligned} \underline{W}^{(\ell)} &:= (W^{(1)}, \dots, W^{(\ell)}) \;, \quad \underline{b}^{(\ell)} := (b^{(1)}, \dots, b^{(\ell)}) \\ W^{(\ell)} &:= W_\ell W_{\ell-1} \cdots W_1 \\ b^{(\ell)} &:= \begin{cases} W_\ell \cdots W_2 b_1 + \cdots + W_\ell b_{\ell-1} + b_\ell & \text{if } \ell \geq 2 \\ b_1 & \text{if } \ell = 1 \,. \end{cases} \end{aligned}$$

Theorem (C-Muñoz Ewald 2023)

The weighted cost function satisfies the upper bound

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

(least square in W_{L+1}, b_{L+1}) where with $X_0 \equiv [\cdots x_{j,i}^{(0)} \cdots] \in \mathbb{R}^{M_0 \times N}$

$$\delta_P := \sup_{j,i} \left| (\overline{(\tau_{\underline{W}^{(L)}, \underline{b}^{(L)}}(X_0))^{red}})^{-1} \Delta(\tau_{\underline{W}^{(L)}, \underline{b}^{(L)}}(x_{j,i}^{(0)})) \right|$$

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

δ_P measures the signal to noise ratio of the truncated training input data.

Strategy to find global cost minimum: Find $\underline{W}^{(L)}, \underline{b}^{(L)}$ so that

$$\Delta(\tau_{\underline{W}^{(L)},\underline{b}^{(L)}}(X_0)) = 0$$

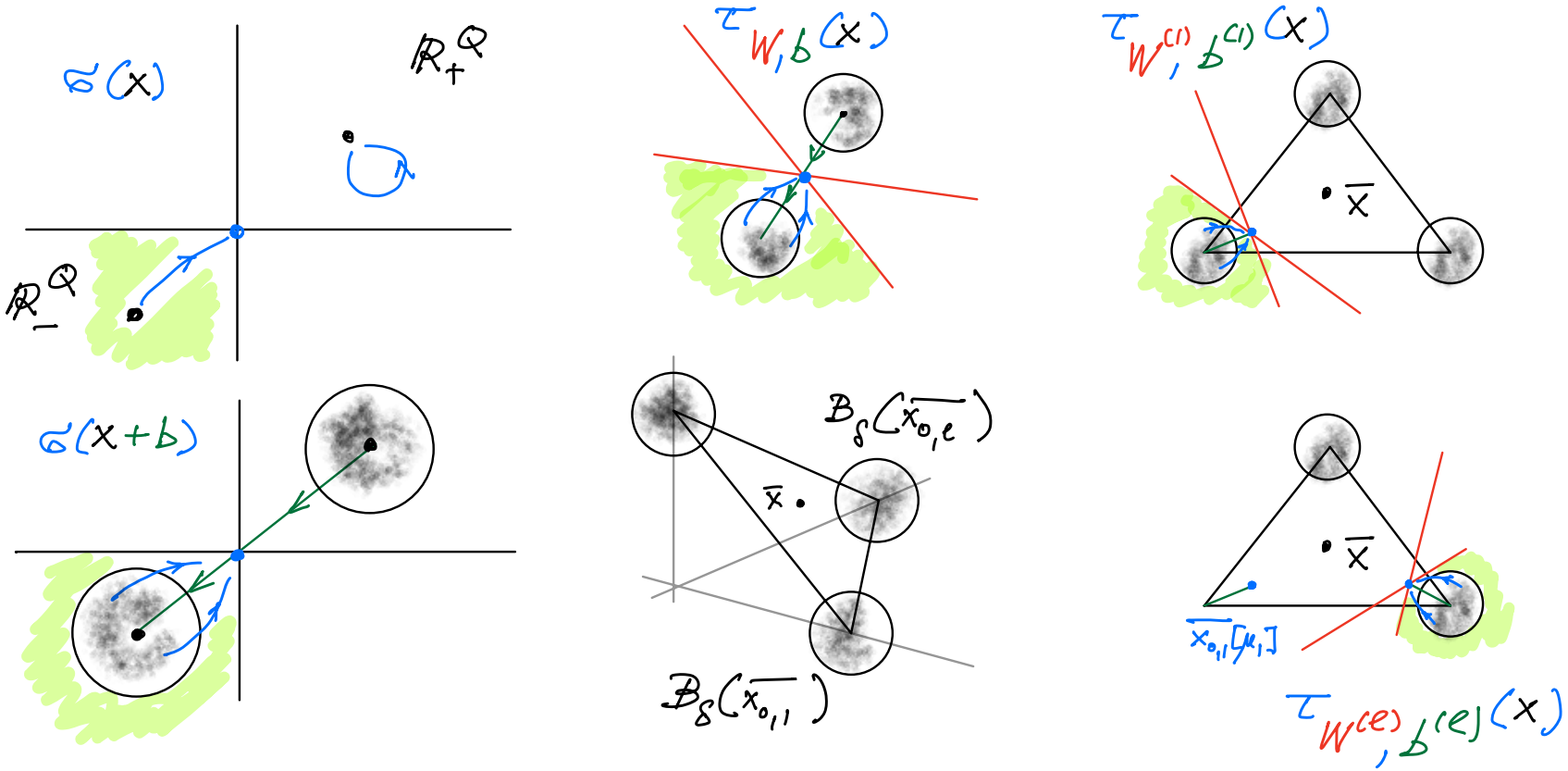
The training inputs belonging to y_j are in δ-ball centered at corner $\overline{x_{0,j}}$ of simplex. Recursively map each of them to a point $\overline{x_{0,j}}[\mu_j]$ on the connecting line from $\overline{x_{0,j}}$ to center of simplex $\overline{x}$, $\mu_j \in \mathcal{I} \subset \mathbb{R}$.

Activation function σ maps positive sector $\mathbb{R}^Q_+$ to itself, and negative sector $\mathbb{R}^Q_-$ to 0. Use $W^{(\ell)}$ to orient diagonal in $\mathbb{R}^Q_+$ from $\overline{x_{0,\ell}}$ towards $\overline{x}$, and use $b^{(\ell)}$ to translate $B_\delta(\overline{x_{0,\ell}})$ into negative sector. Also, choose $W^{(\ell)}$ to change opening angle of $\mathbb{R}^Q_+$ so that all other δ-balls are not affected.

$\Rightarrow$ iterate, each ℓ corresponds to one hidden layer.

Number of parameters: $Q^3 + Q^2 \ll QN$ underparametrized.

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



Theorem (C-Muñoz Ewald 2023)

The global minimum is attained, and is degenerate,

$$\min_{\underline{W}^{(L)}, W_{L+\mathbf{1}}, \underline{b}^{(L)}, b_{L+\mathbf{1}}} \mathcal{C}^{\tau}_{\mathcal{N}}[\underline{W}^{(L)}, W_{L+1}, \underline{b}^{(L)}, b_{L+1}] = 0$$

The minimizers $\underline{W}^{(L)}_*$, $\underline{b}^{(L)}_*[\underline{\mu}]$ *are explicit,* $\underline{\mu} \in \mathcal{I}^Q \subset \mathbb{R}^Q$.

Define metric $d : \mathbb{R}^Q \times \mathbb{R}^Q \to \mathbb{R}_+$ *on the input space,*

$$d(x, x') := \left| Y(\overline{X^{red}_0}[\underline{\mu}])^{-1}(x - x') \right|$$

where $\overline{X^{red}_0}[\underline{\mu}] = [\, \overline{x_{0,1}}[\mu_1] \, \cdots \, \overline{x_{0,Q}}[\mu_Q] \,]$ *and* $Y = [y_1 \ldots y_Q]$.
To match a test input $x \in \mathbb{R}^Q$ *to an output* y_j, *determine*

$$\min_j \; d(\tau_{\underline{W}^{(L)}_*, \underline{b}^{(L)}_*[\underline{\mu}]}(x) \, , \, \overline{x_{0,j}}[\mu_j])$$

Revisit cost function with j-th cluster average in output layer

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

Then, global minimization splits into

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

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.

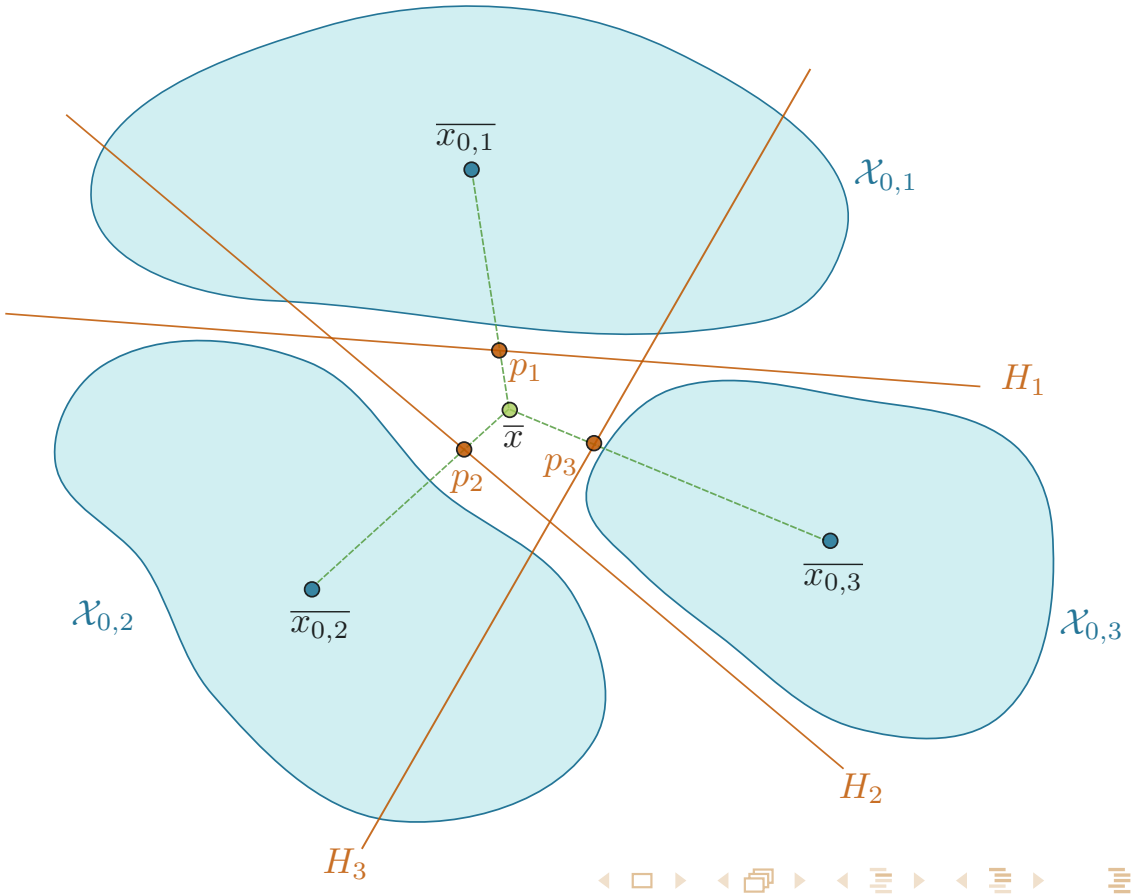
Theorem (C-Muñoz Ewald 2024)

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 minimizers with $Q(M+2)$ parameters.

Unique ordering

$1 \to 2 \to 3$

Sequential application of conical approximation to support vector machine



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

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

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

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

Then, $\mathcal{L}^2$ cost is

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

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

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

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

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

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

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*

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

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 $s \in \mathbb{R}_+$,

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

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*

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

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

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

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)

The standard and the modified gradient flow have the same critical points, and

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

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

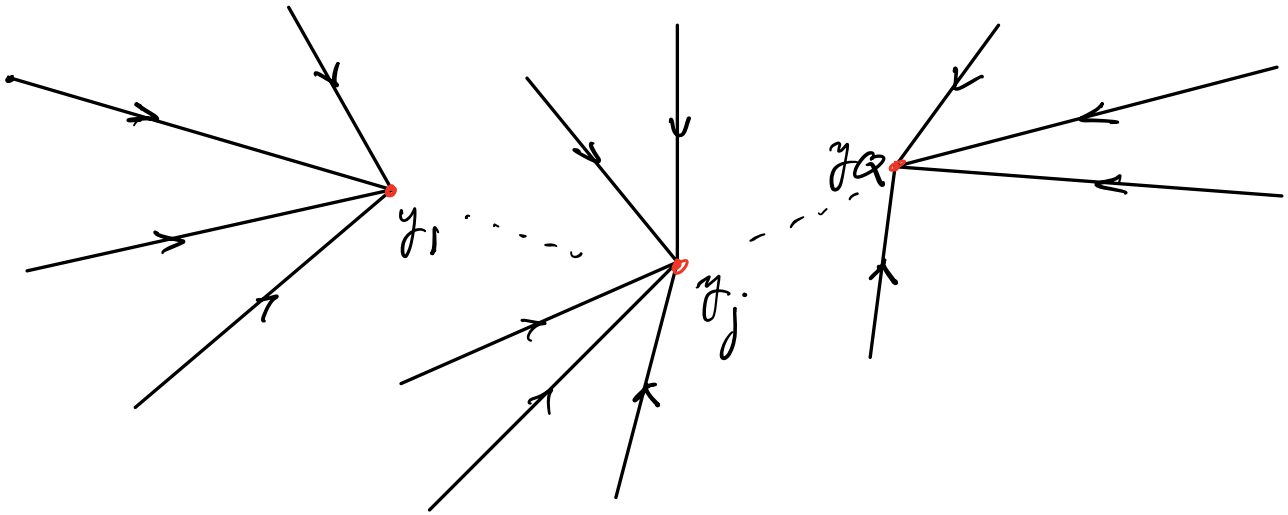
If D *has full rank, then the time reparametrization* $t = 1 - e^{-s/N}$,

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

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

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

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.

Conclusion

Underparametrized DL

Global cost minimization through constructive coarse graining. 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 then follows.

Thank you for your attention !