Explicit construction of global $\mathcal{L}^2$ cost minimizers in underparametrized Deep Learning networks

Thomas Chen

University of Texas at Austin

Includes joint work with

Patricia Muñoz Ewald

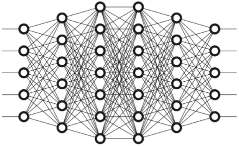
University of Texas at Austin

University of Toronto, 2023

Financial support by NSF.

Deep Learning Networks

DL network for supervised learning: Architecture inspired by brain structure, as designed by nature.



Input layer $X_0 \in \mathbb{R}^{M \times N}$

L hidden layers $X^{(\ell)} \in \mathbb{R}^{M_\ell \times N}$

Output layer $X^{(L+1)} \in \mathbb{R}^{Q \times N}$

Outputs $Y^{ext} \in \mathbb{R}^{Q \times N}$

Parametrized by weight matrices W_ℓ, bias vectors b_ℓ, $\ell = 1, \dots, L+1$

Minimize cost function $\mathcal{C}_\mathcal{N} = \|X^{(L+1)} - Y^{ext}\|^2_{\mathcal{L}^2_\mathcal{N}}$

Definition of DL network

Output matrix

$$Y := [y_1, \ldots, y_Q] \in \mathbb{R}^{Q \times Q}$$

where $y_j \in \mathbb{R}^Q$ is the j-th output vector. Lin indep, invertible.
Training inputs: i-th belonging to y_j.

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

Matrix of all training inputs belonging to y_j

$$X_{0,j} := [x_{0,j,1} \cdots x_{0,j,i} \cdots x_{0,j,N_j}] \, .$$

Matrix of all training inputs, $N := \sum_{j=1}^{Q} N_j$

$$X_0 := [X_{0,1} \cdots X_{0,j} \cdots X_{0,Q}] \in \mathbb{R}^{M \times N}$$

L hidden layers: For $\ell = 1, \dots, L$, recursively define

$$X^{(\ell)} := \sigma(W_\ell X^{(\ell-1)} + B_\ell) \quad \in \mathbb{R}^{M_\ell \times N},$$

Weight matrices

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

Bias vectors $b_\ell \in \mathbb{R}^{M_\ell}$,

$$B_\ell = [b_\ell \cdots b_\ell] \quad \in \mathbb{R}^{M_\ell \times N}$$

Activation function σ (nonlinear !), acting component-wise

$$\begin{aligned} \sigma : \mathbb{R}^{M \times M'} &\to \mathbb{R}_+^{M \times M'} \\ A = [a_{ij}] &\mapsto [(a_{ij})_+] \end{aligned}$$

via ramp function (ReLU)

$$(a)_+ := \max\{0, a\}$$

Terminal layer without activation function, $M_{L+1} = Q$,

$$X^{(L+1)} := W_{L+1}X^{(L)} + B_{L+1} \quad \in \mathbb{R}^{Q\times N}$$

Weighted cost function

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

This is equivalent to Hilbert-Schmidt norm

$$\mathcal{C}_{\mathcal{N}}[(W_i, b_i)_{i=1}^{L+1}] = \left\| X^{(L+1)} - Y^{ext} \right\|_{\mathcal{L}^2_{\mathcal{N}}}^2$$

$$Y^{ext} := [Y_1 \cdots Y_Q] \in \mathbb{R}^{Q\times N} \quad , \quad Y_j := [y_j \cdots y_j] \in \mathbb{R}^{Q\times N_j}$$

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:

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

Let

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

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

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

$\mathcal{C}[\underline{x}[\underline{\theta}(s)]] \geq 0$ 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 are 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.

Construction of global minimizers in underparametrized DL

Joint work with Patricia Muñoz Ewald, 2023.

Assume $M = M_\ell = Q$.

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

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

for $j = 1, \dots, Q$, and

$$\overline{X_{0,j}} := [\overline{x_{0,j}} \cdots \overline{x_{0,j}}] \in \mathbb{R}^{Q \times N_j} \quad , \quad \overline{X_0} := [\overline{X_{0,1}} \cdots \overline{X_{0,Q}}] \in \mathbb{R}^{Q \times N}$$

$$\overline{X_0^{red}} := [\overline{x_{0,1}} \cdots \overline{x_{0,Q}}] \in \mathbb{R}^{Q \times Q}$$

We also define deviations from $\overline{x_{0,j}}$ belonging to output y_j

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

$$\Delta X_{0,j} := [\Delta x_{0,j,1} \cdots \Delta x_{0,j,i} \cdots \Delta x_{0,j,N_j}] \in \mathbb{R}^{Q\times N_j}$$

and total matrix of deviations

$$\Delta X_0 := [\Delta X_{0,1} \cdots \Delta X_{0,j} \cdots \Delta X_{0,Q}] \in \mathbb{R}^{Q\times N}$$

Definition

Given $W \in GL(Q)$, $b \in \mathbb{R}^Q$, and $B = [b \cdots b]$, define the *truncation map*

$$\begin{array}{rcl} \tau_{W,b} : \mathbb{R}^{Q\times N} & \to & \mathbb{R}^{Q\times N} \\ X & \mapsto & W^{-1}(\sigma(WX+B)-B), \end{array}$$

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

We say that $\tau_{W,b}$ is rank preserving with respect to X if both

$$\begin{array}{rcl} \mathrm{rank}(\tau_{W,b}(X)) & = & \mathrm{rank}(X) \\ \mathrm{rank}(\overline{\tau_{W,b}(X)}) & = & \mathrm{rank}(\overline{X}) \end{array}$$

hold, and that it is rank reducing otherwise.

Proposition (C-Muñoz Ewald 2023)

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

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

where (recursive structure similar to renormalization map in RG)

$$\begin{aligned} \tau_{\underline{W}^{(\ell)}, \underline{b}^{(\ell)}}(X_0) &:= \tau_{W^{(\ell)}, b^{(\ell)}}(\tau_{W^{(\ell-1)}, b^{(\ell-1)}}(\cdots \tau_{W^{(2)}, b^{(2)}}(\tau_{W^{(1)}, b^{(1)}}(X_0)) \cdots)) \\ &= \tau_{W^{(\ell)}, b^{(\ell)}}(\tau_{\underline{W}^{(\ell-1)}, \underline{b}^{(\ell-1)}}(X_0)) \end{aligned}$$

$$\begin{aligned} \underline{W}^{(\ell)} &:= (W^{(1)}, \dots, W^{(\ell)}) \ , \ \ \underline{b}^{(\ell)} := (b^{(1)}, \dots, b^{(\ell)}) \\ W^{(\ell)} &:= W_\ell W_{\ell-1} \cdots W_1 \\ b^{(\ell)} &:= \begin{cases} W_\ell \cdots W_2 b_1 + \cdots + W_\ell b_{\ell-1} + b_\ell & \text{if } \ell \geq 2 \\ b_1 & \text{if } \ell = 1 \, . \end{cases} \\ B^{(\ell)} &= [b^{(\ell)} \cdots b^{(\ell)}] \in \mathbb{R}^{Q \times N} \, . \end{aligned}$$

Theorem (C-Muñoz Ewald 2023)

The weighted cost function satisfies the upper bound

$$\min_{\underline{W}^{(L)},W_{L+1},\underline{b}^{(L)},b_{L+1}} \mathcal{C}^{\tau}_{\mathcal{N}}[\underline{W}^{(L)}, W_{L+1}, \underline{b}^{(L)}, b_{L+1}] \leq \quad (1 - C_0\delta_P^2) \min_{\underline{W}^{(L)},\underline{b}^{(L)}} \|Y\,\Delta_1^{(L)}\|_{\mathcal{L}^2_{\mathcal{N}}} \,,$$

(least square in W_{L+1}, b_{L+1}*) for a constant* $C_0 \geq 0$*, where*

$$\Delta_1^{(L)} \quad := \quad (\overline{(\tau_{\underline{W}^{(L)},\underline{b}^{(L)}}(X_0))^{red}})^{-1}\Delta(\tau_{\underline{W}^{(L)},\underline{b}^{(L)}}(X_0))$$

and where

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

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

We note that $\Delta_1^{(0)}$ has the following geometric meaning. Let

$$\Gamma_{\overline{X_0^{red}}} := \left\{ x \in \mathbb{R}^Q \,\middle|\, x = \sum_{j=1}^{Q} \kappa_j \overline{x_{0,j}} \;,\; \kappa_j \geq 0 \,,\; \sum_{j=1}^{Q} \kappa_j = 1 \right\}.$$

Simplex with barycentric coordinates $\kappa = (\kappa_1, \ldots, \kappa_Q)^T \in \mathbb{R}^Q$.
Any point $x \in \mathbb{R}^Q$ can be represented in terms of

$$x = \sum_{i=1}^{Q} \kappa_i \overline{x_{0,i}} = [\overline{x_{0,1}} \cdots \overline{x_{0,Q}}]\kappa = \overline{X_0^{red}}\kappa \,,$$

therefore,

$$\kappa = (\overline{X_0^{red}})^{-1} x$$

are the barycentric coordinates of x. This means that

$$\Delta_1^{(0)} = (\overline{X_0^{red}})^{-1} \Delta X_0$$

is the representation of ΔX_0 in barycentric coordinates !

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

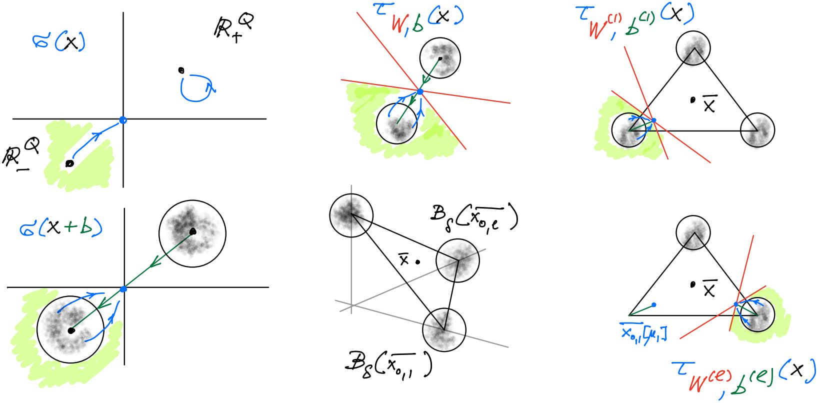
$$\Delta_1^{(L)} = ((\overline{\tau_{\underline{W}^{(L)},\underline{b}^{(L)}}(X_0))^{red}})^{-1} \Delta(\tau_{\underline{W}^{(L)},\underline{b}^{(L)}}(X_0)) = 0$$

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.



$\sigma(x)$
$\mathbb{R}_+^Q$
$\mathbb{R}_-^Q$
$\sigma(x+b)$
$\tau_{W,b}(x)$
$\tau_{W^{(1)},b^{(1)}}(x)$
$\mathcal{B}_\delta(\overline{x_{0,\ell}})$
$\overline{x}$
$\mathcal{B}_\delta(\overline{x_{0,1}})$
$\overline{x}$
$\overline{x_{0,1}}[\mu_1]$
$\tau_{W^{(\ell)},b^{(\ell)}}(x)$

Theorem (C-Muñoz Ewald 2023)

The global minimum is attained, and is degenerate,

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

The minimizers $\underline{W}_*^{(L)}$, $\underline{b}_*^{(L)}[\underline{\mu}]$ *are explicit,* $\underline{\mu} \in \mathcal{I}^Q \subset \mathbb{R}^Q$.
To match a test input $x \in \mathbb{R}^Q$ *to an output* y_j *where* $j = j(x)$

$$\begin{aligned} j(x) &= \mathrm{argmin}_j |W_*^{(L+1)} \tau_{\underline{W}_*^{(L)}, \underline{b}_*^{(L)}[\underline{\mu}]}(x) + b_*^{(L+1)} - y_j| \\ &= \mathrm{argmin}_j \; d(\tau_{\underline{W}_*^{(L)}, \underline{b}_*^{(L)}[\underline{\mu}]}(x) \; , \; \overline{x_{0,j}}[\mu_j]) \end{aligned}$$

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

$$d(x, x') := \left| Y(\overline{X_0^{red}}[\underline{\mu}])^{-1}(x - x') \right|$$

where $\overline{X_0^{red}}[\underline{\mu}] = [\; \overline{x_{0,1}}[\mu_1] \; \cdots \; \overline{x_{0,Q}}[\mu_Q] \;] \in GL(Q)$.

Geometric structure of DL networks

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

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

Def: Comparison model, gradient flow 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 $\underline{x} - \underline{y}_\omega$.

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

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

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

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

Observe that with Jacobian matrix $D[\underline{\theta}]$ for $\underline{x} : \mathbb{R}^K \to \mathbb{R}^{QN}$,

$$\nabla_{\underline{\theta}} \mathcal{C}[\underline{x}[\underline{\theta}]] = D^T[\underline{\theta}] \nabla_{\underline{x}} \mathcal{C}[\underline{x}[\underline{\theta}]] \,.$$

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

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

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

Letting $\underline{x}(s) := \underline{x}[\underline{\theta}(s)]$, so that $\partial_s \underline{x}(s) = -D[\underline{\theta}(s)]\partial_s \underline{\theta}(s)$

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

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" !

Question: Does this make geometric sense ??

Theorem (C 2023)

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*

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

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

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

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

Theorem (C 2023, continued)

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

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

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

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

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

$$\lim_{s \to \infty} \mathcal{C}[\underline{x}[\underline{\theta}(s)]] = 0 \;\; , \;\; \lim_{s \to} \underline{x}[\underline{\theta}(s)] = \underline{y}_{\omega} \, ,$$

at same uniform exponential convergence rates as in comparison model.

Geometric meaning of modified gradient flow

General setting: $\mathcal{M}$ and $\mathcal{N}$ manifolds, $k = \dim(\mathcal{M}) > n = \dim(\mathcal{N})$.
Riemannian structure $(\mathcal{N}, g)$ with metric g.
Smooth surjection $f : \mathcal{M} \to \mathcal{N}$. Pullback map f^*, pushforward map f_*.
$\Gamma(\mathcal{M})$ vector fields (sections) $V : \mathcal{M} \to T\mathcal{M}$.
$\mathcal{V} \subset T\mathcal{M}$ pullback vector bundle $f^*T\mathcal{N}$, with sections $\Gamma(\mathcal{V})$: For $z \in \mathcal{M}$, fiber $\mathcal{V}_z \subset T_z\mathcal{M}$ is spanned by f^*w for $w \in T_{f(z)}\mathcal{N}$.
Define the pullback metric h on $\mathcal{V}$ by way of

$$h(V, W) = g(f_*V, f_*W) \ , \ \text{for } V, W \in \Gamma(\mathcal{V})$$

Define the gradient grad_h associated to $(\mathcal{M}, \mathcal{V}, h)$ by way of

$$d\mathcal{F}(V) = h(V, \mathrm{grad}_h \mathcal{F}) \ , \ \text{for all } V \in \Gamma(\mathcal{V})$$

any smooth $\mathcal{F} : \mathcal{M} \to \mathbb{R}$, with exterior derivative $d\mathcal{F}$.

In local coordinates,

$$h(V,W) = g_{\alpha,\alpha'} Df^{\alpha}_{\beta} Df^{\alpha'}_{\beta'} V^{\beta} W^{\beta'} \,,$$

and for all $V \in \Gamma(\mathcal{V})$, with $d\mathcal{F}(V) = V^{\beta}\partial_{\beta}\mathcal{F}$,

$$V^{\beta}\partial_{\beta}\mathcal{F} = g_{\alpha,\alpha'} Df^{\alpha}_{\beta} Df^{\alpha'}_{\beta'} V^{\beta} (\mathrm{grad}_h \mathcal{F})^{\beta'}$$

In our DL situation, $\mathcal{M} = \mathbb{R}^K$, $\mathcal{N} = \mathbb{R}^{QN}$ so that $k = K$ and $n = QN$.
f corresponds to $\underline{x} : \mathbb{R}^K \to \mathbb{R}^{QN}$, $\underline{\theta} \mapsto \underline{x}[\underline{\theta}]$.
Pushforward f_* with Jacobian matrix $[Df^{\alpha}_{\beta}] = D[\underline{\theta}]$ at $\underline{\theta} \in \mathbb{R}^K$.
Fiber $\mathcal{V}_{\underline{\theta}}$ of pullback vector bundle $\mathcal{V}$ is the range of $D^T[\underline{\theta}]$.
Riemannian structure on $\mathcal{N} = \mathbb{R}^{QN}$ $\Leftrightarrow$ Euclidean metric, $g_{\alpha,\alpha'} = \delta_{\alpha,\alpha'}$

$$\underline{V}^T \nabla_{\underline{\theta}} \mathcal{F} = \underline{V}^T D^T[\underline{\theta}] D[\underline{\theta}] \, \mathrm{grad}_h \mathcal{F} \,,$$

for all $\underline{V} = (V^1, \dots, V^K)^T \in \mathrm{range}(D^T[\underline{\theta}])$. Equivalent to

$$P[\underline{\theta}] \nabla_{\underline{\theta}} \mathcal{F} = P[\underline{\theta}] D^T[\underline{\theta}] D[\underline{\theta}] \, \mathrm{grad}_h \mathcal{F}$$

where $P = P^2 = P^T \in \mathbb{R}^{K \times K}$ orthoprojector onto $\mathrm{range}(D^T)$.

Applying the Penrose inverse of $D^T[\underline{\theta}]$ from the left,

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

subsequently applying the Penrose inverse of $D[\underline{\theta}]$ from the left,

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

$P\mathrm{grad}_h\mathcal{F} = \mathrm{grad}_h\mathcal{F}$ and $P^{\perp}\mathrm{grad}_h\mathcal{F} = 0 \Rightarrow \mathrm{grad}_h\mathcal{F} \in \Gamma(\mathcal{V})$ section of $\mathcal{V}$
We conclude that, writing $\widetilde{\mathcal{C}}[\underline{\theta}] := \mathcal{C}[\underline{x}[\underline{\theta}]]$ for the $\mathcal{L}^2$ cost function,

$$\partial_s\underline{\theta}(s) = -\mathrm{grad}_h\widetilde{\mathcal{C}}[\underline{\theta}(s)]$$

If $\mathcal{V}$ non-integrable (non-holonomic), triple

$$(\mathbb{R}^K, \mathcal{V}, h)$$

defines a *sub-Riemannian manifold* with grad_h on $\mathcal{V}$.

Thank you for your attention !