A Tropical Geometry Perspective on Learning from Data: Challenges and Opportunities

Santiago VELASCO-FORERO

PSL Research University / École des Mines de Paris

September 18, 2025

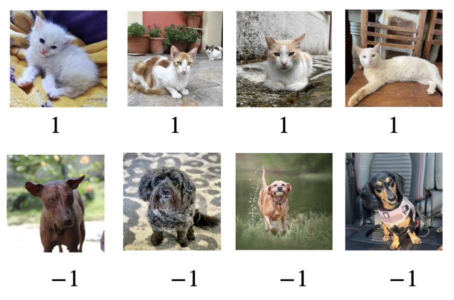
Outline

Learning from Data

- Given N observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \ldots, N$ (**Data**),
- The objective is to find a function $f_\theta : \mathcal{X} \to \mathcal{Y}$ (**model**) with $\theta \in \mathbb{R}^d$ (**Parameters**)
- to correct predict the observation $x \in \mathcal{X}$ (**Training data**)
- to correct predict a new previously unseen $x^{\texttt{new}} \in \mathcal{X}$ (**Testing data**)

Learning from Data

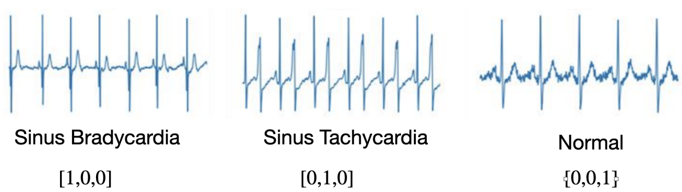
RGB images ($H \times W$ pixels)



$\mathcal{X} \in \mathcal{F}(\mathbb{Z}^{H \times W}, \mathbb{R}^3)$
$\mathcal{Y} = \{-1, 1\}$

Learning from Data

1D Signal of length M.



$\mathcal{X} \in \mathcal{F}(\mathbb{Z}^M, \mathbb{R})$
$\mathcal{Y}$ is the probability simplex.

Learning from Data

3D Point Cloud (M points)

$\mathcal{X} \in \mathcal{F}(\mathbb{R}^{M\times 3}, \mathbb{R})$
$\mathcal{Y}$ is the probability simplex.

Learning from Data

- Given N observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \ldots, N$ (**Data**),
- The objective is to find a function $f_\theta : \mathcal{X} \to \mathcal{Y}$ (**model**) with $\theta \in \mathbb{R}^d$ (**Parameters**)
- to correct predict the observation $x \in \mathcal{X}$ (**Training data**)
- to correct predict a new previously unseen $x^{\texttt{new}} \in \mathcal{X}$ (**Testing data**)

Risk of a model

The **risk** associated with the model f_θ is defined as the expectation of the loss function $\texttt{loss} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$, i.e,

$$\texttt{risk}(f_\theta) = \int \texttt{loss}(f_\theta(x), y) dP(x, y)$$

Learning from Data

- Given N observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \dots, N$ (**Data**), **(i.i.d)**
- The objective is to find a function $f_\theta : \mathcal{X} \to \mathcal{Y}$ (**model**) with $\theta \in \mathbb{R}^d$ (**Parameters**)
- to correct predict the observation $x \in \mathcal{X}$ (**Training data**)
- to correct predict a new previously unseen $x^{\texttt{new}} \in \mathcal{X}$ (**Testing data**)

Empirical Risk of a model

The **empirical risk** associated with the model f_θ is defined as the average of the loss function on training data

$$\texttt{risk}_{\texttt{emp}}(f_\theta) = \frac{1}{N} \sum_{i=1}^{N} \texttt{loss}(f_\theta(x_i), y_i)$$

Learning from Data

ERM principle

The **empirical risk minimization principle** states that the learning algorithm should choose a model f_θ^* which minimize the empirical risk over the model class $\mathcal{H}$:

$$f_\theta^* = \arg\min_{f \in \mathcal{H}} \mathtt{risk}_{\mathtt{emp}}(f_\theta) \quad (1)$$

Learning from Data

ERM principle

The **empirical risk minimization principle** states that the learning algorithm should choose a model f_θ^* which minimize the empirical risk over the model class $\mathcal{H}$:

$$f_\theta^* = \arg\min_{f \in \mathcal{H}} \mathtt{risk}_{\mathtt{emp}}(f_\theta) \tag{2}$$

The two main questions are:

1. Which family of functions are we going to optimize?
2. How do we perform the optimization?

Learning from Data

empirical risk minimization principle

The **empirical risk minimization principle** states that the learning algorithm should choose a model f_θ^* which minimize the empirical risk over the model class $\mathcal{H}$:

$$f_\theta^* = \arg\min_{f \in \mathcal{H}} \mathtt{risk}_{\mathtt{emp}}(f_\theta) + \lambda\Omega(\theta)$$

The two main questions are:

1. How do we perform the optimization? **(Not in this talk)**
2. Which family of functions are we going to optimize?

Nowdays Approach

1. **Static Models**: They are composed of linear functions $f_{\theta_i} := f_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i+1}}$ called **layers** with nonlinear **activation functions** applied componentwise to all the layers.

$$x \xrightarrow{f_0} \ldots \xrightarrow{f_i} \ldots \xrightarrow{f_r} y$$

   - universal approximators (in the sense that they are dense in L^2).
   - they do not have many guaranteed properties besides continuity.

2. **Dynamic models**

Nowdays Approach

1. **Dynamic models**[1]

$$\begin{array}{ccccc} & x & \xrightarrow{f_0} \dots \xrightarrow{f_i} \dots \xrightarrow{f_K} & x_r & \\ g_0 \downarrow & & \downarrow g_i & & \downarrow g_r \\ & y_0 & y_i & y_r & \end{array}$$

- ARIMA models
- Recurrent neural networks
- Long short-term memory
- Diffusion models

[1] Algebraic Dynamical Systems in Machine Learning, I. Jones et al., 2024, Applied Categorical Structures

Static Models

- Neural Networks: $f_\theta = \theta_r^T \sigma(\theta_{r-1}^T \sigma(\ldots \theta_2^T \sigma(\theta_1^T x)))$

Withdraws:

1. Non-convex optimization problems.
2. Generalization guarantees in the overparameterized regime.
3. Energy consuming in both training and inference.
4. Too big to fail?

Adversarial Examples : Non-Lipschitz functions

- Given a model $f_{\boldsymbol{\theta}}$ and *a small perturbation* δ, we call $\mathbf{x}^{adv}$ an adversarial example if there exists $\mathbf{x}$, an example drawn from the benign data distribution, such that $||f_{\boldsymbol{\theta}}(\mathbf{x}) - f_{\boldsymbol{\theta}}(\mathbf{x}^{adv})|| > \delta$ and $||\mathbf{x} - \mathbf{x}^{adv}|| \leq \epsilon$.
- An human user would still visually consider the adversarial input $\mathbf{x}^{adv}$ similar to or the same as the benign input x
- Usually, we are interested in adversarial examples for benign samples $\mathbf{x}$, i.e., samples where the model gives a correct prediction.

Non-Lipschitz functions

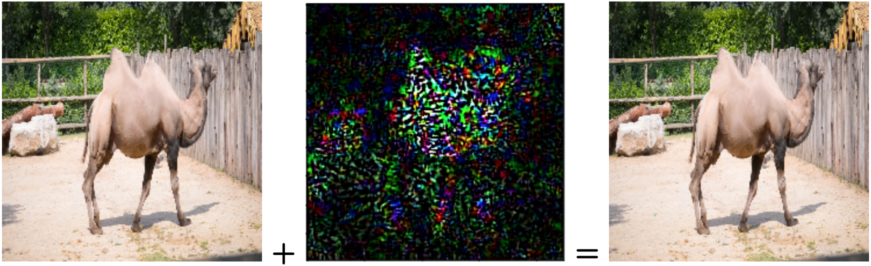


Figure: $\mathbf{x} + \epsilon = \mathbf{x}^{adv}$. For a CNN, the prediction in $\mathbf{x}$ is a **Camel**, but for $\mathbf{x}^{adv}$ is a **dog**

Non symmetries

'Egyptian_cat', 0.3396838

'lynx', 0.47152225

'jay', 0.96423554

VGG19

'plastic_bag', 0.54238236

'electric_ray', 0.8287997

Include invariances

- Translation Invariances $\rightarrow$ Convolutional version
- Symmetries $\rightarrow$ Group CNNs
- Other geometries?

Geometric Deep Learning

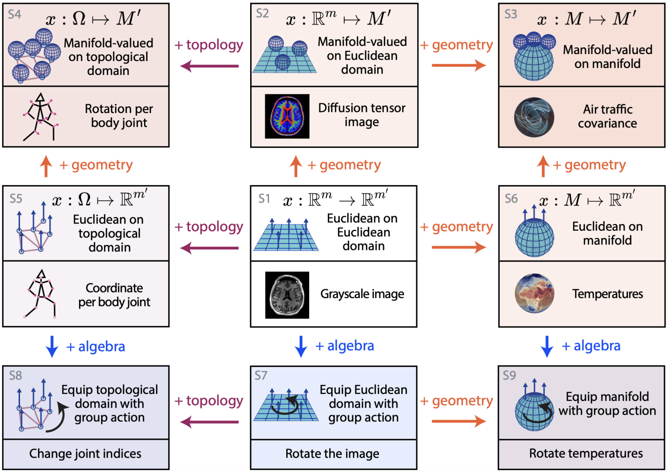


Figure: Beyond Euclid: An Illustrated Guide to Modern Machine Learning with Geometric, Topological, and Algebraic Structures, M. Papillon et al., 2025

Tropical Version

The adjective “tropical” was coined by French mathematicians Dominique Perrin and Jean-Eric Pin, to honor their Brazilian colleague Imre Simon, a pioneer of min-plus algebra as applied to finite automata in computer science.

Tropical geometry is a marriage between algebraic geometry and polyhedral geometry. A piecewise-linear version of algebraic geometry. [Maclagan and Sturmfels 2015]

Tropical Semifield

$\mathbb{R}_{max} = \mathbb{R} \cup \{-\infty\}$ equipped with
$a + b = max(a, b)$ and $a \times b = a + b$,
$0 = -\infty$ $1 = 0$
Dual semifield: $\mathbb{R}_{min} = \mathbb{R} \cup \{+\infty\}$
equipped with $a + b = min(a, b)$, instead of max.

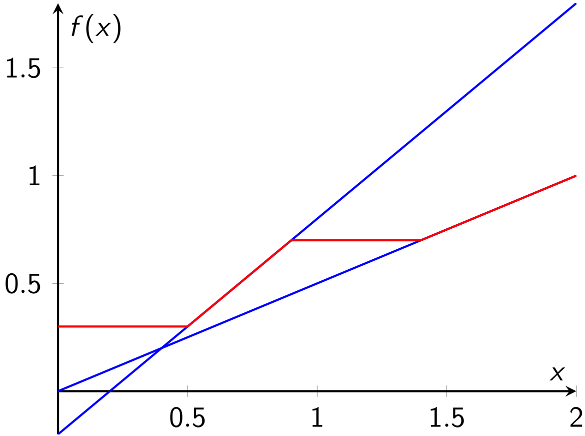


Figure: f(x)=min(max(x-0.2,0.3),max(x/2,0.7))

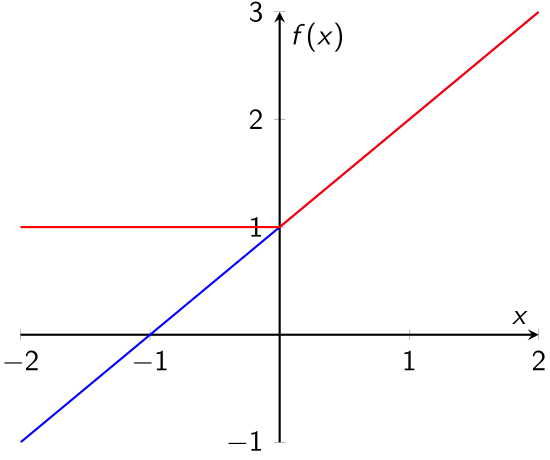


Figure: The tropical line $f(x) = max(x+1, 1)$

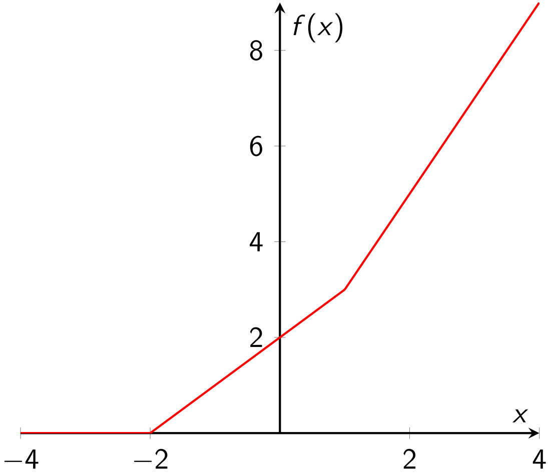


Figure: The tropical parabola $f(x) = max(1 + 2x, 2 + x, 0)$

Tropical Approach

1. **Hybrid** Static Models: They are composed of linear function followed by **tropical functions** $f_{\theta_i} := f_i : \mathbb{R}^{d_i} \rightarrow \mathbb{R}^{d_{i+1}}$.

$$x \xrightarrow{f_0} \ldots \xrightarrow{f_i} \ldots \xrightarrow{f_r} y$$

   - universal approximators?

Theorem ([Gorokhovik et al., 1994][Bartels et al., 1995][Ovchinnikov, 2002])

Let f be a PL function on a closed convex domain $\Omega \subset \mathbb{R}^n$ and $\{g_1 = \beta_1 x + \alpha_1, \cdots, g_d = \beta_d x + \alpha_d\}$ be the set of the d linear components of f, with $\beta_i, \alpha_i \in \mathbb{R}^n$. There is a family $\{K_i\}_{i \in I}$ of subsets of set $\{1, \cdots, d\}$ such that

$$f(x) = \max_{i \in I} \min_{j \in K_i} g_j(x), \;\; x \in \Omega. \tag{3}$$

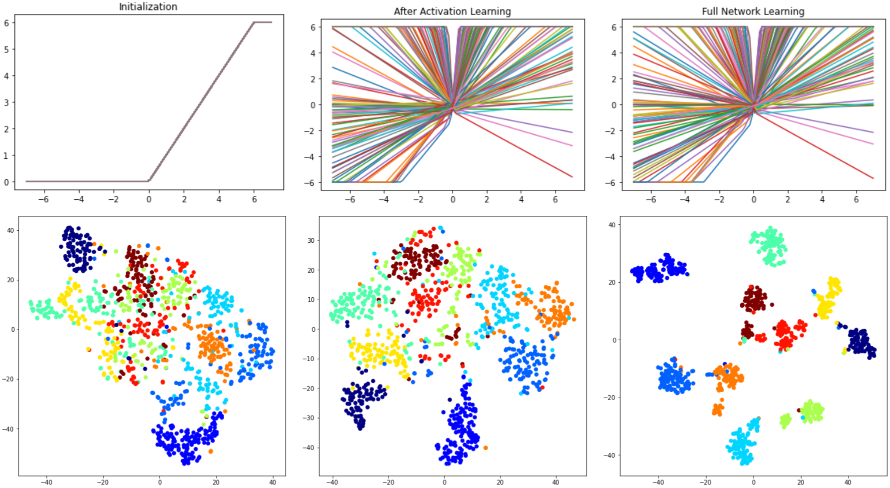


Figure: First Row: Left: Random Initialisation with (**14%**) of accuracy. We use a simplified version of proposed activation $\min(\max(\beta_0 x + \alpha_0, \beta_1 x + \alpha_1, \alpha_2), \alpha_3)$, with initialisation $\max(\min(\max(x, 0), 6), -6)$ Centre: Training only activation functions (**92.38%**), Right: Training Full Network (**98,58%**). Second Row: t-SNE visualisation of last layer is the 10-classes MNIST prediction for a CNN.

Include invariances

- Translation Invariances $\rightarrow$ Convolutional version $\rightarrow$ **Sup-convolutions**
- Symmetries $\rightarrow$ Group CNNs $\rightarrow$ **Group Morphology** [2]
- Other geometries? $\rightarrow$ **Working in progress**

[2] V. Penaud–Polge et al. Group Equivariant Morphological Networks, SIAM JOIS, 2025 (Accepted)

Sup convolution

We study here functions $f : E \to \overline{\mathbb{R}}$, where $\overline{\mathbb{R}}$ it allowed to be *extended-real-valued*, *i.e.*, to take values in $\overline{\mathbb{R}} = [-\infty, \infty]$. Accordingly, the set of all such functions is denoted by $\mathcal{F}(E, \overline{\mathbb{R}})$.

Definition

The **sup-convolution** $\delta_\theta(f)$ of f is defined by:

$$\delta_\theta(f)(x) := \sup_{y \in E} \{f(y) + \theta(x - y)\} = \sup_{w \in E} \{f(x - w) + \theta(w)\} \quad (4)$$

where $\theta \in \mathcal{F}(E, \overline{\mathbb{R}})$ is the (additive) structuring function which determines the effect of the operator. Here the inf-addition rule $\infty - \infty = \infty$ is to be used in case of conflicting infinities. $\sup f$ and $\inf f$ refer to the *supremum* (least upper bound) and *infimum* (greatest lower bound) of f. In the discrete case where the function is a finite set of points, max and min are used.

Inf convolution

Definition

The **inf-convolution** $\varepsilon_\theta(f)$, is the adjoint operator to the sup-convolution 4, and it is defined as

$$\varepsilon_\theta(f)(x) := -\delta_{\check{\theta}}(-f)(x) = \inf_{y \in E} \{f(y) - \theta(y - x)\} = \inf_{w \in E} \{f(x + w) - \theta(w)\} \quad (5)$$

where the transposed structuring function is $\check{\theta}(x) = \theta(-x)$.

$\forall f, g \in \mathcal{F}(E, \overline{\mathbb{R}})$

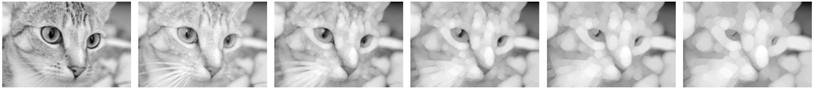
1. The operators (4) and (5) are translation invariant.
2. (4) and (5) correspond to one another through the duality relation $\delta_\theta(f)(x) \leq g(x) \iff f(x) \leq \varepsilon_\theta(g)(x)$, called **adjunction** or **Galois connection**.
3. An operator ξ is called *increasing* if $f(x) \geq g(x) \Rightarrow \xi(f)(x) \geq \xi(g)(x)$ $\forall x$. The sup-conv (4) and inf-conv (5) are increasing for all θ.
4. An operator ξ is called *extensive* (resp. *antiextensive*) if $\xi(f)(x) \geq f(x)$ (resp. $\xi(f)(x) \leq f(x)$), $\forall x$. The sup-conv (4) (resp. erosion (5)) is extensive (resp. antiextensive) if and only if $\theta(0) \geq 0$, *i.e.*, the structuring function evaluated at the origin is non-negative.
5. $\varepsilon_\theta(f)(x) \leq f(x) \leq \delta_\theta(f)(x)$ if and only if $\theta(0) \geq 0$.
6. δ_θ (resp. ε_θ) does not introduce any local maxima (resp. local minima) if $\theta \leq 0$ and $\theta(0) = 0$. In this case, we say that θ is *centered*.

Theorem (Maragos (1989))

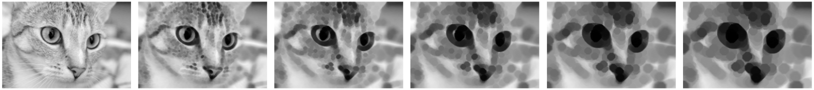
Consider an upper semi-continuous operator Ψ *acting on an upper semi-continuous function. Let* $Bas(\Psi) = \{g_i\}_{i \in I}$ *be its basis and* $Bas(\bar{\Psi}) = \{h_j\}_{j \in J}$ *the basis of the dual operator. If* Ψ *is a TI and increasing operator then it can be represented as*

$$\begin{aligned} \Psi(f)(x) &= \sup_{i \in I} (f \ominus g_i)(x) = \sup_{i \in I} \inf_{y \in \mathbb{R}^n} \{f(x+y) - g_i(y)\} && (6) \\ &= \inf_{j \in J} (f \oplus \check{h}_j)(x) = \inf_{j \in J} \sup_{y \in \mathbb{R}^n} \left\{f(x-y) + \check{h}_j(y)\right\} && (7) \end{aligned}$$

Example of Max-Plus convolution by iterating



Example of Min-Plus convolution by iterating



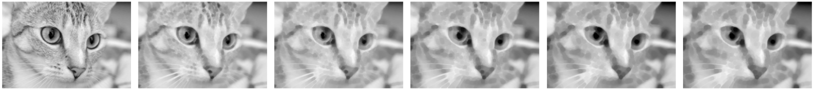
Example of plus-times convolution by iterating

Closure and Kernel Operator

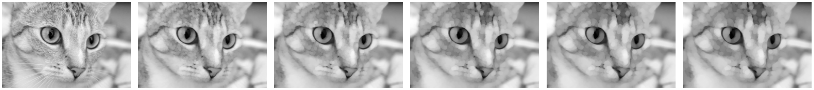
Definition

Given a Galois connection with lower adjoint F and upper adjoint G, we can consider the compositions $G \circ F$, known as the associated **closure operator**, and $F \circ G$, known as the associated **kernel operator**. Both are monotone and idempotent, and we have $f \leq G \circ F(f)$ for all f in A and $F \circ G(f) \leq f$ for all b in B.

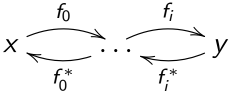
Example of Closure Operator by changing scale parameter



Example of Kernel Operator by changing scale parameter



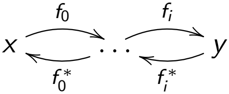
Assume a static model composed of tropical functions $f_{\theta_i} := f_i$. Then the static model is increasing and extensive (or antiextensive). Additionally, the adjoint operator give a closed-form for f_i^*k



3

[3] T. Leeuwen et al, , An invertible generative model for forward and inverse problems, 2025.

Assume a static model composed of tropical functions $f_{\theta_i} := f_i$. Then the static model is increasing and extensive (or antiextensive). Additionally, the adjoint operator give a closed-form for f_i^*k

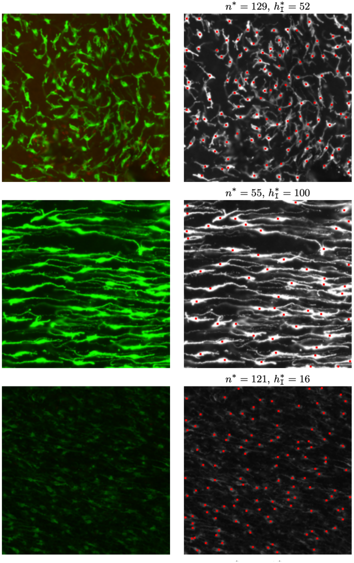


[3] We can do something that cannot be done with plus-times convolutions?.

[3]T. Leeuwen et al, , An invertible generative model for forward and inverse problems, 2025.


Three cats and two dogs
Enter a negative prompt
Generate image

We can learn to count!



Reconstruction[4]

Definition

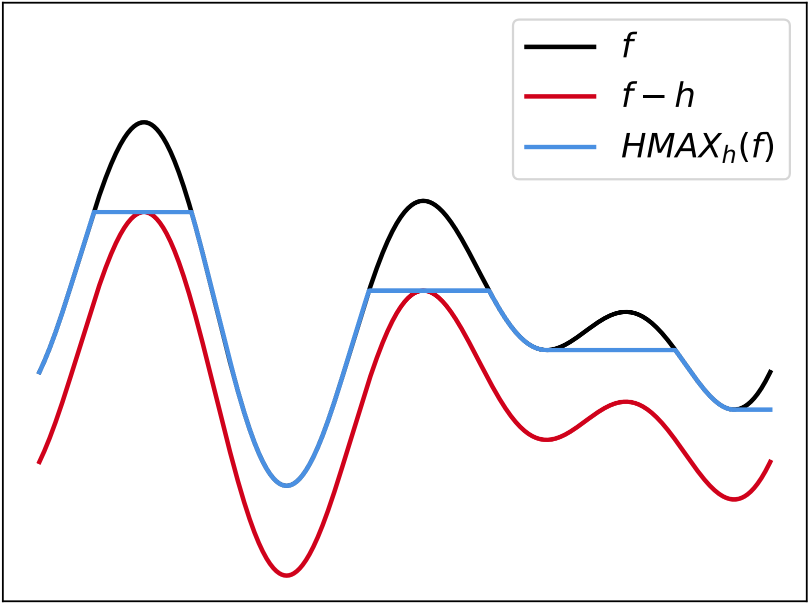
$\forall f, g \in \mathcal{F}(E, \overline{\mathbb{R}})$, the reconstruction of f from g is defined as :

$$\mathtt{REC}(f,g)(x) = \max_{y \in \Omega, \gamma \in \Gamma_{xy}} \left(f(y) \wedge \min_{z \in \gamma} g(z) \right). \tag{8}$$

where Γ_{xy} denotes the set of path between x and y.

Note that $\mathtt{REC}(f,g)(x)$ is increasing and antiextensive operator.

[4] Blusseau, S. et al(2025). Cell counting with trainable h-maxima and connected component layers. JMIV 67(3), 1-27.

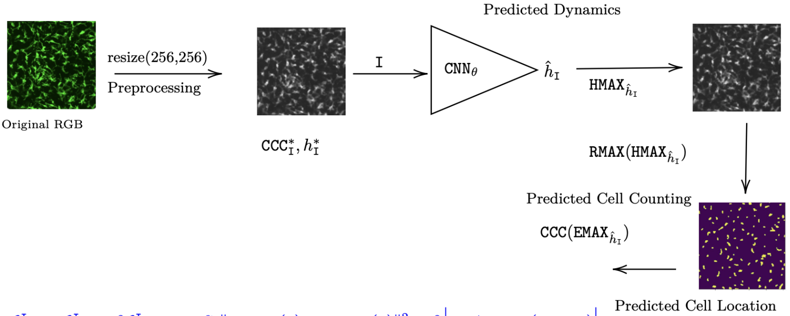


(a)

Example of Reconstruction by Max-plus with different parameters of dynamic

Example of Reconstruction by Min-plus with different parameters of dynamic

Example of Reconstruction by Max-plus with different parameters of dynamic



$$\mathcal{L}^{\mathtt{I}} = \alpha\mathcal{L}^{\mathtt{I}}_{rec} + \beta\mathcal{L}^{\mathtt{I}}_{count} := \frac{\alpha}{|\Omega|}\|\mathtt{HMAX}_{h^*_{\mathtt{I}}}(\mathtt{I}) - \mathtt{HMAX}_{\hat{h}_{\mathtt{I}}}(\mathtt{I})\|_2^2 + \beta\left|\mathtt{CCC}^*_{\mathtt{I}} - \mathtt{CCC}(\mathtt{EMAX}_{\hat{h}_{\mathtt{I}}})\right|$$

Dataset	Method	#Param	$\mathcal{A}_{err}$(%)	$\mathcal{T}_{err}$(%)	MAE	MPE(%)
TRP1 [3]	Lazard *et al* [3]	1,760,000	**9.28**	**8.72**	-	-
	Joint loss, (MB,-1)	16,675	12.6 ±0.6	10.4 ±0.5	**5.82** ±0.3	**-5.47** ±0.8
	Joint loss, (MB,N=50)	16,675	13.6 ±1.2	11.1 ±0.8	6.21 ±0.45	-7.15 ±1.8
	Count. loss (MB, -1)	16,675	13.7 ±1.3	11.0 ±1.1	6.2 ±0.6	-8.60 ±1.8
	Count. loss (MB, N=50)	16,675	12.9 ±0.36	10.4 ±0.38	5.84 ±0.2	-7.76 ±0.81
Fluorescent Neuronal Cells [4]	Morelli *et al* [4]	888,977	-	-	3.09	**-5.13**
	Joint loss (MB, -1)	16,675	34.4 ±2.6	28.6 ±0.4	2.89 ±0.04	-9.1 ±5.6
	Joint loss (MB, N=50)	16,675	33.0 ±2.1	28.1 ±0.7	2.84 ±0.07	-6.05 ±3.5
	Count. loss (MB, -1)	16,675	**31.7** ±1.2	28.1 ±0.7	2.84 ±0.08	7.12 ±1.6
	Count. loss (MB, N=50)	16,675	32.1 ±1.2	**25.3** ±0.8	**2.56** ±0.08	-7.17 ±3.3
Cellpose [20]	Unet [22]	7,852,033	12.1 ±2.1	11.8 ±1.9	6.31 ±1.0	11.2 ±2.5
	Joint loss (MB, -1)	16,675	**6.98** ±0.7	8.09 ±0.6	4.34 ±0.33	**0.25** ±1.4
	Joint loss (MB, N=50)	16,675	7.01 ±0.93	7.97 ±0.82	4.28 ±0.44	1.35 ±1.34
	Count. loss (MB, -1)	16,675	7.10 ±1.2	**7.47** ±1.7	**4.01** ±0.89	0.94 ±1.2
	Count. loss (MB, N=50)	16,675	8.87 ±1.2	10.3 ±1.3	5.52 ±0.7	4.75 ±2.3

Thanks!

Collaborators:

1. Valentin Penaud–Polge
2. Mihaela Dimitrova
3. Samy Blusseau
4. Gustavo Angulo
5. Xiahu Liu
6. Marco Valle (Campinas University)

ANR: Deep Ordering for Vector-Valued Operators and Neural Networks – DEEPORDER 

Bartels, S. G., Kuntz, L., and Scholtes, S. (1995).
Continuous selections of linear functions and nonsmooth critical point theory.
Nonlinear Analysis: Theory, Methods & Applications, 24(3):385–407.

Gorokhovik, V. V., Zorko, O. I., and Birkhoff, G. (1994).
Piecewise affine functions and polyhedral sets.
Optimization, 31(3):209–221.

Ovchinnikov, S. (2002).
Max–min representations of piecewise linear functions.
Beiträge Algebra Geom., 43:297–302.