EPFL

Supervised Learning Over Banach Spaces

Shayan Aziznejad
Biomedical Imaging Group
EPFL, Lausanne, Switzerland

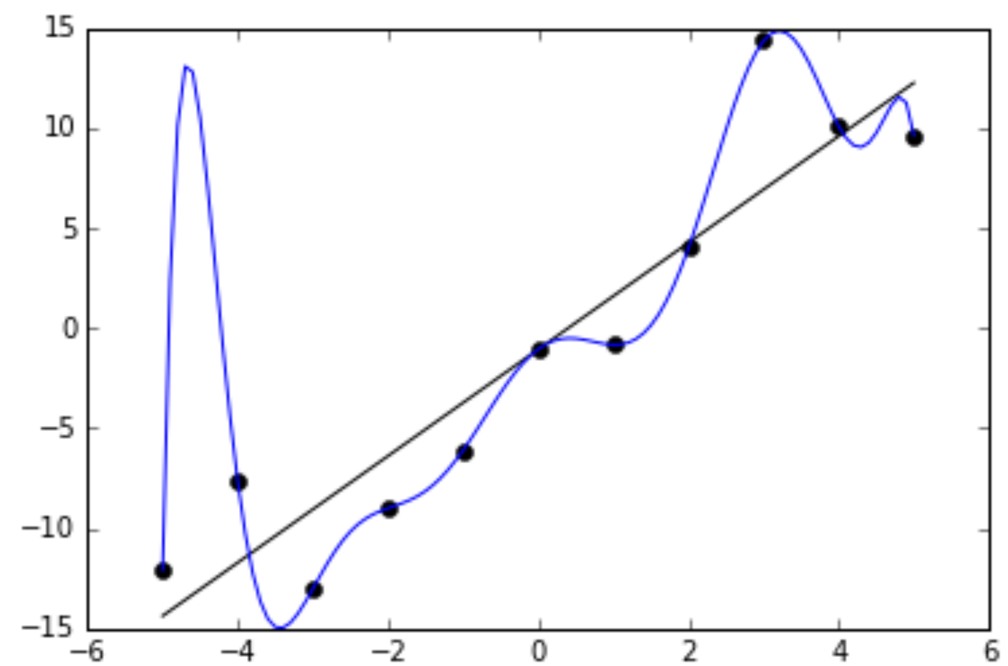
November 3, 2021

Supervised Learning

- Training Data: $(\boldsymbol{x}_m, y_m) \subseteq \mathbb{R}^d \times \mathbb{R}$ for $m = 1, \ldots, M$

- Goal: Find $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(\boldsymbol{x}_m) \approx y_m$ for all m

Without $\overbrace{\text{Overfitting!}}^{\text{malignant}}$



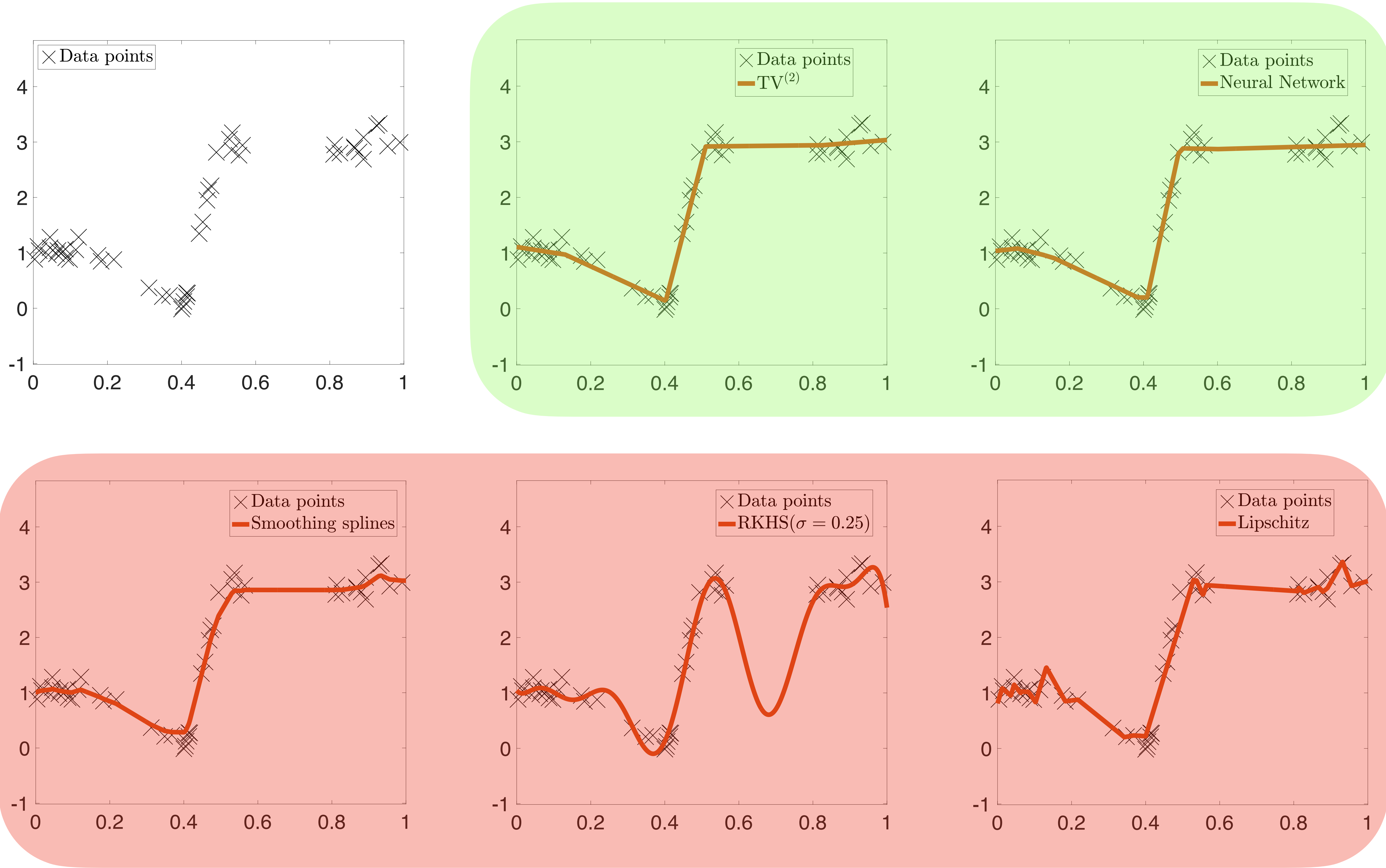
Source: en.wikipedia.org/wiki/Overfitting

Variational Formulation of Learning

$$\min_{f \in \mathcal{F}(\mathbb{R}^d)} \underbrace{\sum_{m=1}^{M} E(f(\boldsymbol{x}_m), y_m)}_{\text{Data Fidelity}} + \underbrace{\lambda \mathcal{R}(f)}_{\text{Regularization}}$$

- $\mathcal{F}(\mathbb{R}^d)$: Search space
  - Parametric regression: *e.g.* Neural networks with a prescribed architecture
  - Nonparametric regression: *e.g.* Reproducing kernel Hilbert space (RKHS)
- $E : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$: Convex loss function
  - *e.g.* Quadratic loss $E(y, z) = (y - z)^2$
- $\mathcal{R} : \mathcal{F}(\mathbb{R}^d) \to \mathbb{R}_{\geq 0}$: Regularization functional
  - Weight decay in deep learning
  - The squared RKHS norm

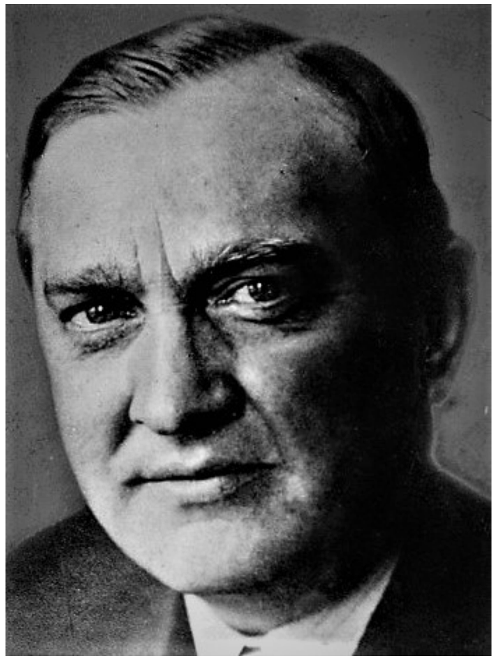
Example


Data points
Data points
$TV^{(2)}$
Data points
Neural Network
Data points
Smoothing splines
Data points
RKHS($\sigma = 0.25$)
Data points
Lipschitz

OUTLINE

- **Introduction** ✔
- **Learning over Banach spaces**
  - Theory of Banach spaces
  - General representer theorem
  - Application: Sparse multikernel regression
- **Learning activation functions of DNNs**
  - One-dimensional learning
  - Deep splines
- **Going to higher dimensions**
  - Hessian-based regularization
- **Future works**

Banach Spaces



Stefan Banach (1892 – 1945)

- $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$: Complete normed vector space
  - Strong topology: $x_k \to x$ if $\|x_k - x\|_{\mathcal{X}} \to 0$
- Finite-dimensional examples
  - $\left(\mathbb{R}^N, \|\cdot\|_p\right)$, where $\|\mathbf{a}\|_p = \begin{cases} \left(\sum_{n=1}^N |a_n|^p\right)^{\frac{1}{p}}, & p \in [1, +\infty) \\ \max_n |a_n|, & p = +\infty \end{cases}$
  - $\left(\mathbb{R}^{M\times N}, \|\cdot\|_{S_p}\right)$, where $\|\mathbf{A}\|_{S_p} = \|\boldsymbol{\sigma}(\mathbf{A})\|_p$ (Schatten-p Norm)
- Infinite-dimensional examples
  - $\left(L_p(\mathbb{R}^d), \|\cdot\|_{L_p}\right)$, where $\|f\|_{L_p} = \begin{cases} \left(\int_{\mathbb{R}^d} |f(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}\right)^{\frac{1}{p}}, & p \in [1, +\infty) \\ \mathrm{ess\ sup}_{\boldsymbol{x}\in\mathbb{R}^d} |f(\boldsymbol{x})|, & p = +\infty \end{cases}$
  - $\left(\mathcal{C}_0(\mathbb{R}^d), \|\cdot\|_{L_\infty}\right)$: Continuous functions that vanish at infinity

Dual of a Banach Space

- $(\mathcal{X}', \|\cdot\|_{\mathcal{X}'})$: Space of continuous linear functionals $\mathcal{X} \to \mathbb{R}$
  - $x' : x \mapsto x'(x) = \langle x', x \rangle_{\mathcal{X}' \times \mathcal{X}} = \langle x', x \rangle$
  - $\|x'\|_{\mathcal{X}'} = \sup_{\|x\|_{\mathcal{X}}=1} \langle x', x \rangle$
- Examples $\quad p \in [1, +\infty]$ and $q = \frac{p}{p-1}$
  - $\left(\mathbb{R}^N, \|\cdot\|_p\right)' = \left(\mathbb{R}^N, \|\cdot\|_q\right)$
  - $\left(\mathbb{R}^{M\times N}, \|\cdot\|_{S_p}\right)' = \left(\mathbb{R}^{M\times N}, \|\cdot\|_{S_q}\right)$
  - $\left(L_p(\mathbb{R}^d), \|\cdot\|_{L_p}\right)' = \left(L_q(\mathbb{R}^d), \|\cdot\|_{L_q}\right)$ for $p \neq +\infty$
- $\left(\mathcal{C}_0(\mathbb{R}^d), \|\cdot\|_{L_\infty}\right)' = \left(\mathcal{M}(\mathbb{R}^d), \|\cdot\|_{\mathcal{M}}\right)$ (Duval-Peyré '15) (Chizat-Bach '20)
  - Theorem[Riesz-Markov]: $\mathcal{M}(\mathbb{R}^d)$ is the space of finite signed measures

Weak*-Topology and Existence

- $(x_n') \subseteq \mathcal{X}'$ converges in weak*-topology to $x' \in \mathcal{X}'$, if

$$\langle x_n', x \rangle \to \langle x', x \rangle, \quad \forall x \in \mathcal{X}$$

- Theorem[Banach-Alaoglu]: $B_{\mathcal{X}'} = \{\|x'\|_{\mathcal{X}'} \leq 1\}$ is weak*-compact.

- Consequence: Generalized Weierstrass theorem

  - $\mathcal{J} : \mathcal{X}' \to \mathbb{R}_{\geq 0}$: weak*-lower semicontinuous

    $\Rightarrow \arg\min_{\|x'\|_{\mathcal{X}'} \leq C} \mathcal{J}(x')$ is nonempty

  - $\mathcal{J} : \mathcal{X}' \to \mathbb{R}_{\geq 0}$: weak*-lower semicontinuous and coercive

    $\Rightarrow \arg\min_{x' \in \mathcal{X}'} \mathcal{J}(x')$ is nonempty

Duality Mapping and Extreme Points

- Recall: $\|x'\|_{\mathcal{X}'} = \sup_{\|x\|_{\mathcal{X}}=1} \langle x', x \rangle$

- Generic duality bound: $\langle x', x \rangle \leq \|x'\|_{\mathcal{X}'} \|x\|_{\mathcal{X}}$

- Duality mapping: $\mathcal{J}_{\mathcal{X}} : \mathcal{X} \to 2^{\mathcal{X}'}$ (Beurling-Livingston '62)

  - $x' \in \mathcal{J}_{\mathcal{X}}(x)$ if $\|x'\|_{\mathcal{X}'} = \|x\|_{\mathcal{X}}$ and $\langle x', x \rangle = \|x'\|_{\mathcal{X}'} \|x\|_{\mathcal{X}}$

- $\mathcal{J}_{\mathcal{X}}(x) \neq \emptyset$ for all $x \in \mathcal{X}$

- $\mathrm{Ext}(B)$: Extreme point of the convex set B

  - $x \in \mathrm{Ext}(B)$ if $\nexists x_1, x_2 \in B, \alpha \in (0,1) : x = \alpha x_1 + (1-\alpha) x_2$

General Representer Theorem

Theorem [Unser '21, Unser-A.'22]

- $\mathcal{X}'(\mathbb{R}^d)$: Banach space of functions $\mathbb{R}^d \to \mathbb{R}$
- $\boldsymbol{x}_m \in \mathbb{R}^d, m = 1, \dots, M$: distinct data points
- $\forall m, \delta_{\boldsymbol{x}_m} : \mathcal{X}'(\mathbb{R}^d) \to \mathbb{R} : f \mapsto f(\boldsymbol{x}_m)$: weak*-continuous
- $E : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$: Strictly convex

Then, the solution set

$$\mathcal{V} = \arg\min_{f \in \mathcal{X}'(\mathbb{R}^d)} \sum_{m=1}^{M} E\left(f(\boldsymbol{x}_m), y_m\right) + \lambda \|f\|_{\mathcal{X}'}$$

is nonempty, convex and weak*-compact. Moreover:

1. $\exists \nu = \sum_{m=1}^{M} c_m \delta_{\boldsymbol{x}_m} \in \mathcal{X}$ such that $\mathcal{V} \subseteq \mathcal{J}_{\mathcal{X}}(\nu)$
2. $\mathrm{Ext}(\mathcal{V})$: linear combination of at most M extreme points of $B_{\mathcal{X}'}$ (Boyer *et al.* '19)

Example: Hilbert Spaces

David Hilbert
(1862 – 1943)

- $\mathcal{H}(\mathbb{R}^d)$: Complete inner-product space
  - Banach space: $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f\rangle}$
  - Riesz map: Linear isometry $\mathrm{R}_{\mathcal{H}} : \mathcal{H}(\mathbb{R}^d) \to \mathcal{H}'(\mathbb{R}^d)$ with

$$\langle \mathrm{R}_{\mathcal{H}}(f), g\rangle_{\mathcal{H}'\times\mathcal{H}} = \langle f, g\rangle, \qquad \forall f, g \in \mathcal{H}(\mathbb{R}^d)$$

- $\mathcal{H}'(\mathbb{R}^d)$: RKHS ⇔ Weak*-continuity of pointwise evaluation
  - Reproducing kernel: $\mathrm{K}(\cdot, \boldsymbol{x}) = \mathrm{R}_{\mathcal{H}}(\delta_{\boldsymbol{x}})$ for all $\boldsymbol{x} \in \mathbb{R}^d$ (Aronszajn '62)

- Duality mapping: $\mathcal{J}_{\mathcal{H}}(f) = \{\mathrm{R}_{\mathcal{H}}(f)\}$

$$\Rightarrow f^* = \mathrm{R}_{\mathcal{H}}\left(\sum_{m=1}^{M} c_m \delta_{\boldsymbol{x}_m}\right) = \sum_{m=1}^{M} c_m \mathrm{K}(\cdot, \boldsymbol{x}_m)$$

Unique solution

(Scholkopf *et al.* '01) (Wahba '90)

Banach Kernels

- Recall: $\mathcal{M}(\mathbb{R}^d)$ is the space of finite Radon measures
  - $L_1(\mathbb{R}^d) \subseteq \mathcal{M}(\mathbb{R}^d)$ with $\|f\|_{L_1} = \|f\|_{\mathcal{M}}$ for any $f \in L_1(\mathbb{R}^d)$.
  - For any $\boldsymbol{a} = (a_n) \in \ell_1(\mathbb{Z})$:

$$w_{\boldsymbol{a}} = \sum_{n\in\mathbb{Z}} a_n \delta_{\boldsymbol{x}_n} \in \mathcal{M}(\mathbb{R}^d), \qquad \|w_{\boldsymbol{a}}\|_{\mathcal{M}} = \|\boldsymbol{a}\|_{\ell_1}$$

Johann Radon (1887 – 1956)

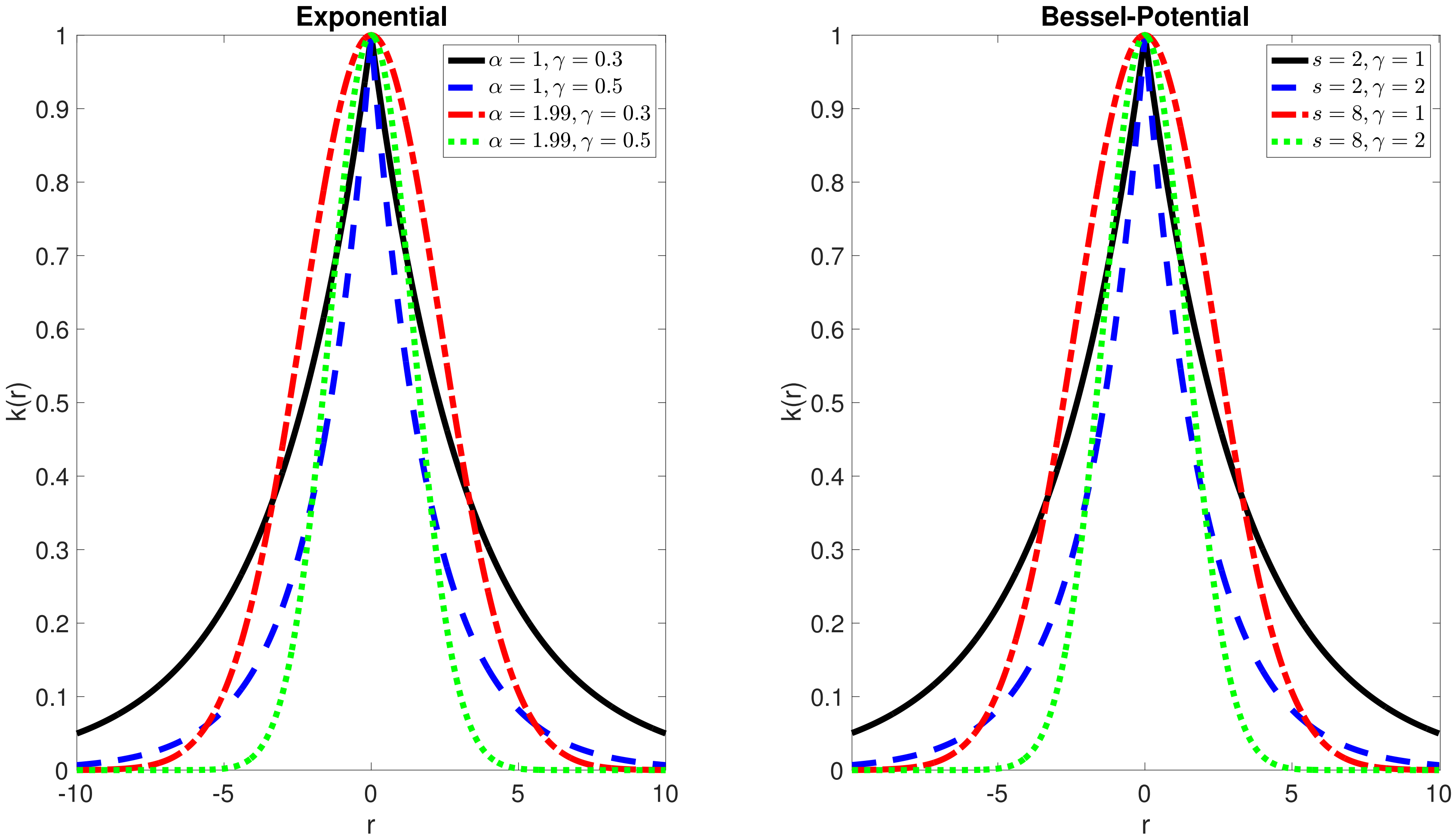
- L: Linear shift-invariant (LSI) isomorphisms onto $\mathcal{M}(\mathbb{R}^d)$
- Search space $\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) = \mathrm{L}^{-1}\left(\mathcal{M}(\mathbb{R}^d)\right)$
  - Banach structure: $\|f\|_{\mathcal{M}_{\mathrm{L}}} = \|\mathrm{L}\{f\}\|_{\mathcal{M}}$
  - Banach kernel: $\mathrm{k} = \mathrm{L}^{-1}\{\delta\} \in \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$

Admissible Banach Kernels

Theorem [A.-Unser '21]

1. The LSI operator L is an isomorphism onto $\mathcal{M}(\mathbb{R}^d)$ if and only if the Fourier transform of its Banach kernel $\widehat{k}(\boldsymbol{\omega})$ is a smooth, nonvanishing, slowly growing, and heavy-tailed function of $\boldsymbol{\omega}$.

2. Pointwise evaluation is weak*-continuous over $\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$, if and only if $\mathrm{k} \in \mathcal{C}_0(\mathbb{R}^d)$.



Sparse Multikernel Regression

- Learning with multiple kernels (Lanckriet *et al.* '04) (Bach *et al.* '05)
  - $\mathrm{k}_1, \dots, \mathrm{k}_N$: prescribed positive-definite kernels
  - Learn a positive-definite kernel $\mathrm{k}_{\boldsymbol{\mu}} = \sum_{n=1}^{N} \mu_n \mathrm{k}_n$ from the data
- Multicomponent model: $f = f_1 + \cdots + f_N, \qquad \forall n : f_n \in \mathcal{M}_{\mathrm{L}_n}(\mathbb{R}^d)$
- Search space: $\mathcal{X}'(\mathbb{R}^d) = \prod_{n=1}^{N} \mathcal{M}_{\mathrm{L}_n}(\mathbb{R}^d)$
  - $\|\mathbf{f}\|_{\mathcal{X}'} = \big\| \big(\|f_n\|_{\mathcal{M}_{\mathrm{L}_n}} \big) \big\|_1 = \sum_{n=1}^{N} \|f_n\|_{\mathcal{M}_{\mathrm{L}_n}}.$
- Extreme points of $B_{\mathcal{X}'}$ [Unser-A. '22]

$$\mathbf{f} = (f_n) \in \mathrm{Ext}\,(B_{\mathcal{X}'}) \Leftrightarrow \exists n_0 \text{ and } \boldsymbol{z} \in \mathbb{R}^d : \mathbf{f} = (0, \dots, \pm \mathrm{k}_{n_0}(\cdot - \boldsymbol{z}), \dots, 0)$$

Sparse Multikernel Regression

Theorem [A.-Unser '21] There exists f^* solution of

$$\min_{\substack{f_n \in \mathcal{M}_{\mathrm{L}_n}(\mathbb{R}^d), \\ f = \sum_{n=1}^N f_n}} \sum_{m=1}^{M} \mathrm{E}(f(\boldsymbol{x}_m), y_m) + \lambda \|(f_n)\|_{\mathcal{X}'},$$

with the expansion

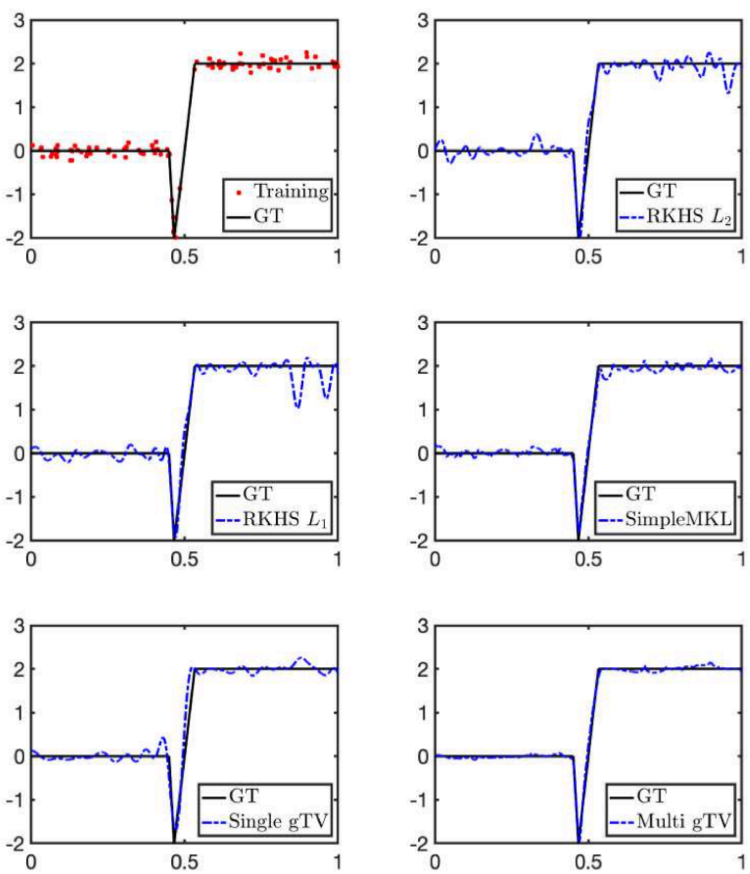
$$f^* = \sum_{n=1}^{N} \sum_{l=1}^{M_n} a_{n,l} \mathrm{k}_n(\cdot, \boldsymbol{z}_{n,l}),$$

where $K = \sum_{n=1}^N M_n \leq M$. Moreover, the unknown kernel coefficients $\boldsymbol{a} = (a_{n,l}) \in \mathbb{R}^K$ are in the solution set of

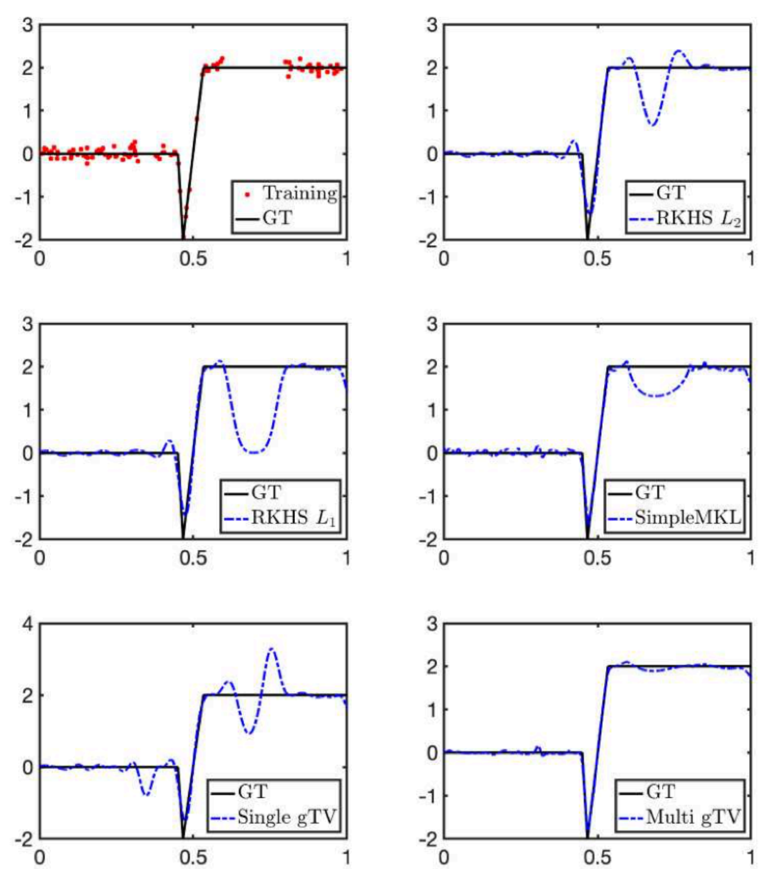
$$\min_{\boldsymbol{a} \in \mathbb{R}^K} \left(\sum_{m=1}^{M} \mathrm{E}([\mathbf{G}\boldsymbol{a}]_m, y_m) + \lambda \|\boldsymbol{a}\|_{\ell_1} \right)$$

for some matrix $\mathbf{G} \in \mathbb{R}^{M \times K}$ that depends on the kernel locations $\boldsymbol{z}_{n,l}$.

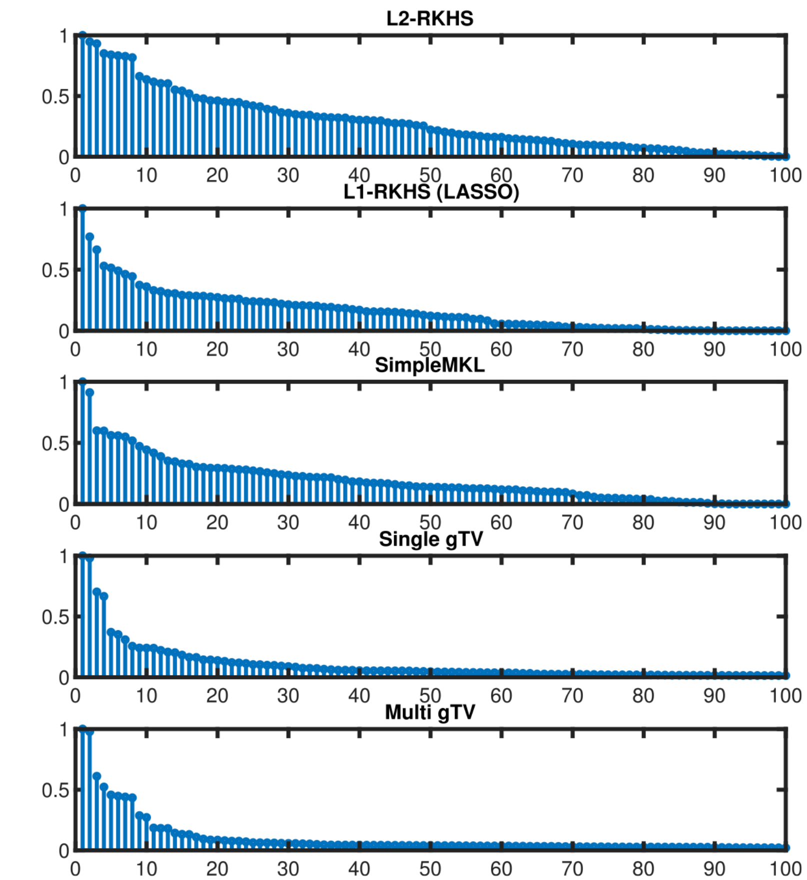
Sparse Multikernel Regression



(a) Full data



(b) Missing data



Quantity	Dataset	L2-RKHS	L1-RKHS	SimpleMKL	Single gTV	Multi gTV
Sparsity	Full data	64.7	44.1	54.4	32.5	**20.0**
	Missing data	66.1	39.3	56.0	32.9	**31.1**
MSE (dB)	Full data	-17.2	-16.1	-15.2	-16.7	**-18.1**
	Missing data	-2.6	-2.7	-10.9	-3.9	**-17.3**

Related Literature

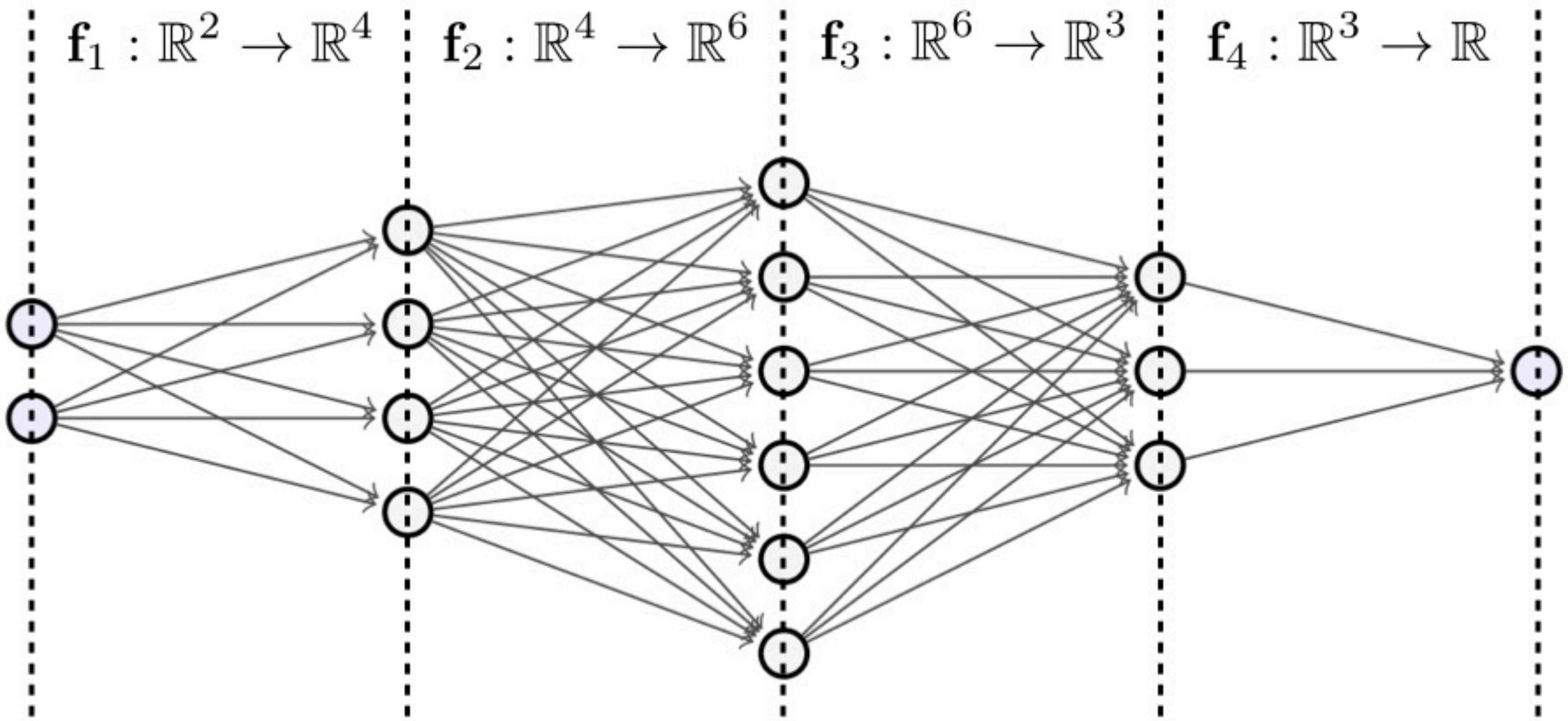
- S.D. Fisher, J.W. Jerome, "Spline solutions to L1 extremal problems in one and several variables," *Journal of Approximation Theory*, 1975.
- E. Mammen, S. van de Geer. "Locally adaptive regression splines," *The Annals of Statistics,* 1997*.*
- M. Unser, J. Fageot, J.P. Ward, "Splines are universal solutions of linear inverse problems with generalized TV regularization," *SIAM Review*, 2017.
- A. Flinth, P. Weiss. "Exact solutions of infinite dimensional total-variation regularized problems," *Information and Inference: A Journal of the IMA,* 2019.
- V. Duval, G. Peyré, "Exact support recovery for sparse spikes deconvolution," *Foundations of Computational Mathematics,* 2015.
- E. Candès, C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on pure and applied Mathematics,* 2014.
- C. Boyer, A. Chambolle, Y. De Castro, V. Duval, F. De Gournay, P. Weiss. "On representer theorems and convex regularization." *SIAM Journal on Optimization* 29, no. 2 (2019): 1260-1281.

OUTLINE

- **Introduction** ✔
- **Learning over Banach spaces** ✔
  - Theory of Banach spaces
  - General representer theorem
  - Application: Sparse multikernel regression
- **Learning activation functions of DNNs**
  - One-dimensional learning
  - Deep splines
- **Going to higher dimensions**
  - Hessian-based regularization
- **Future works**

Deep Neural Networks (DNNs)

- Composition of "simple" vector-valued mappings



$$\mathbf{f}_{\text{deep}} = \mathbf{f}_4 \circ \mathbf{f}_3 \circ \mathbf{f}_2 \circ \mathbf{f}_1 : \mathbb{R}^2 \to \mathbb{R}$$

Feed-Forward DNNs

- Input-output relation $$\mathbf{f}_{\text{deep}} : \mathbb{R}^{N_0} \rightarrow \mathbb{R}^{N_L} : \boldsymbol{x} \mapsto \mathbf{f}_L \circ \cdots \circ \mathbf{f}_1(\boldsymbol{x}).$$

- lth layer $$\mathbf{f}_l(\boldsymbol{x}) = \Big(\sigma_{1,l}(\mathbf{w}_{1,l}^T \boldsymbol{x}), \sigma_{2,l}(\mathbf{w}_{2,l}^T \boldsymbol{x}), \ldots, \sigma_{N_l,l}(\mathbf{w}_{N_l,l}^T \boldsymbol{x})\Big)$$

  - Linear layer $$\mathbf{W}_l = \begin{bmatrix} \mathbf{w}_{1,l} & \mathbf{w}_{2,l} & \cdots & \mathbf{w}_{N_l,l} \end{bmatrix}^T$$

  - Pointwise nonlinearity

$$\boldsymbol{\sigma}_l : \mathbb{R}^{N_l} \rightarrow \mathbb{R}^{N_l} \qquad (x_1, \ldots, x_{N_l}) \mapsto (\sigma_{1,l}(x_1), \sigma_{2,l}(x_2), \ldots, \sigma_{N_l,l}(x_{N_l}))$$

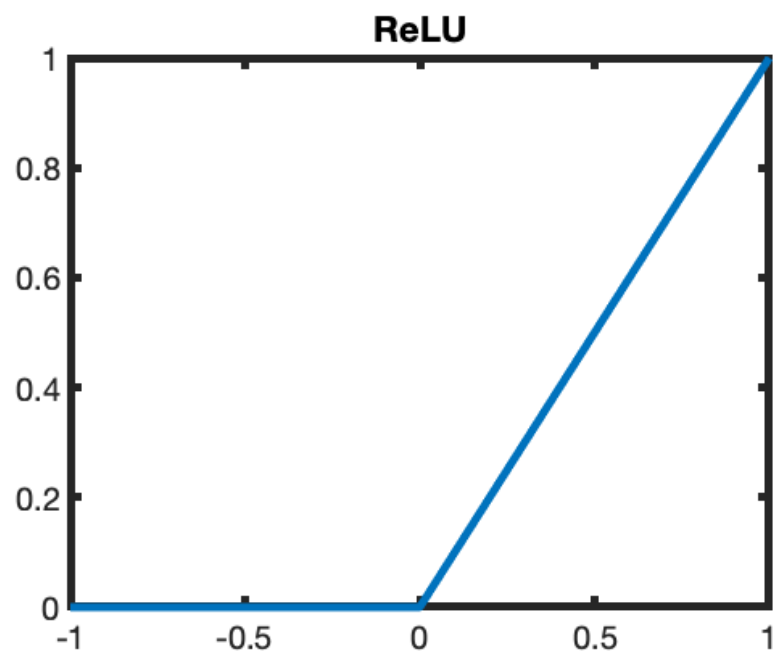
- Alternative representation $$\mathbf{f}_l = \boldsymbol{\sigma}_l \circ \mathbf{W}_l$$

Fixed Activation Functions: ReLU, LReLU

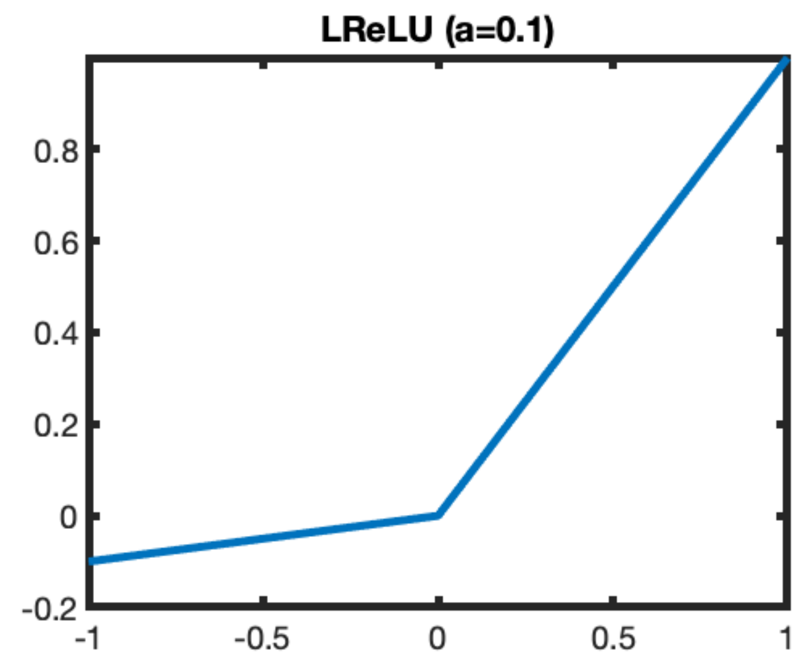
- Fixed-shape Nonlinearities $\sigma_{n,l}(x) = \sigma(x - b_{n,l})$

- Common choices:

$$\mathrm{ReLU}(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad \mathrm{LReLU}_a(x) = \begin{cases} x, & x \geq 0 \\ ax, & x < 0 \end{cases}$$



(Glorot *et al.* '11)



(Maas *et al.* '13)

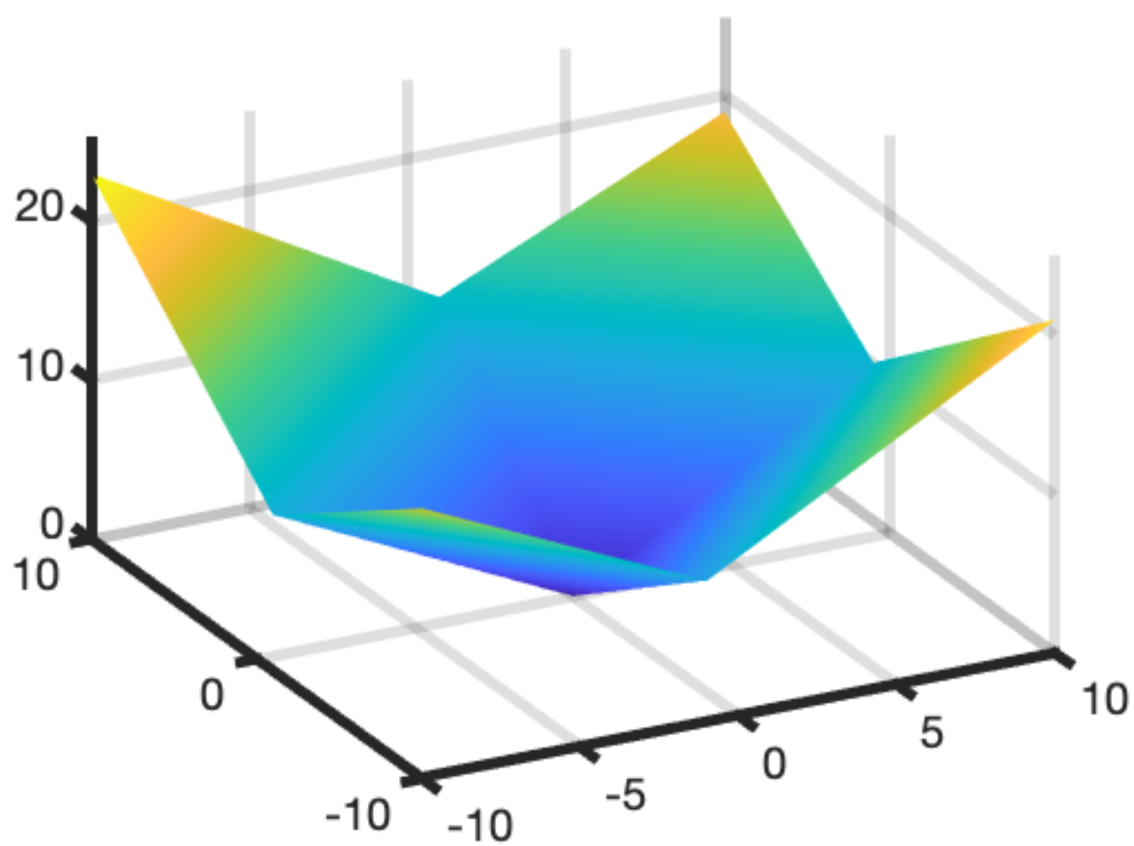
- ReLU DNNs: Hierarchical splines (Poggio *et al.* '15)

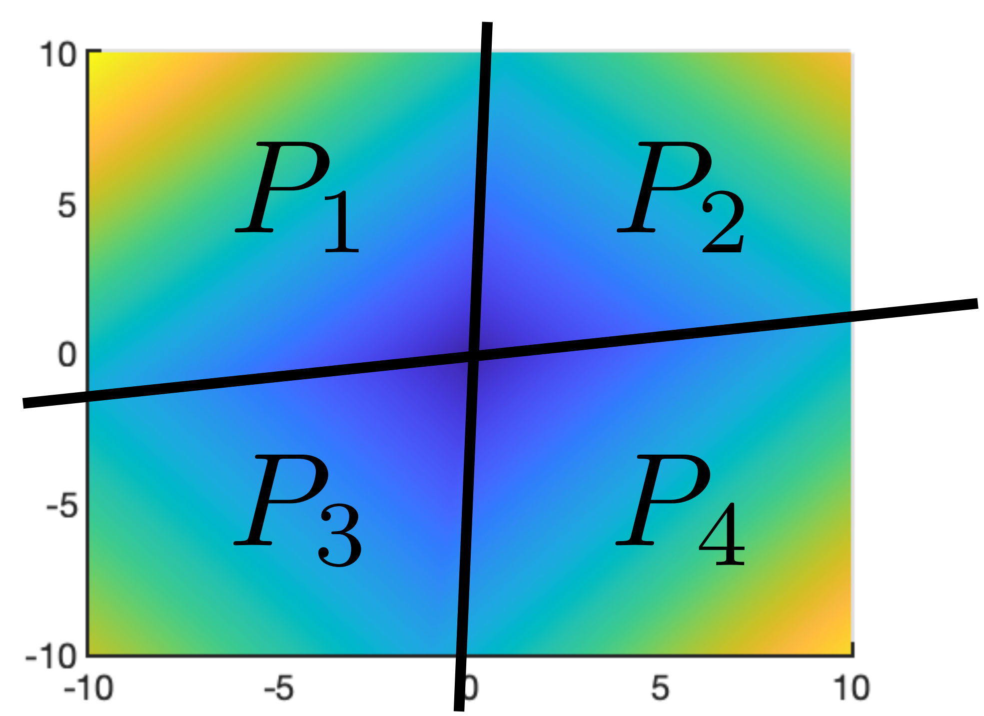
CPWL Structure of DNNs

- Definition (Wang-Sun 2005)

A function $f : \mathbb{R}^{N_0} \to \mathbb{R}$ is continuous piecewise-linear (CPWL) if:

- it is continuous, and,
- its domain $\mathbb{R}^{N_0} = \bigcup_{k=1}^{K} P_k$ can be partitioned into a finite set of non-overlapping convex polytopes P_k over which it is affine.





CPWL Structure of DNNs

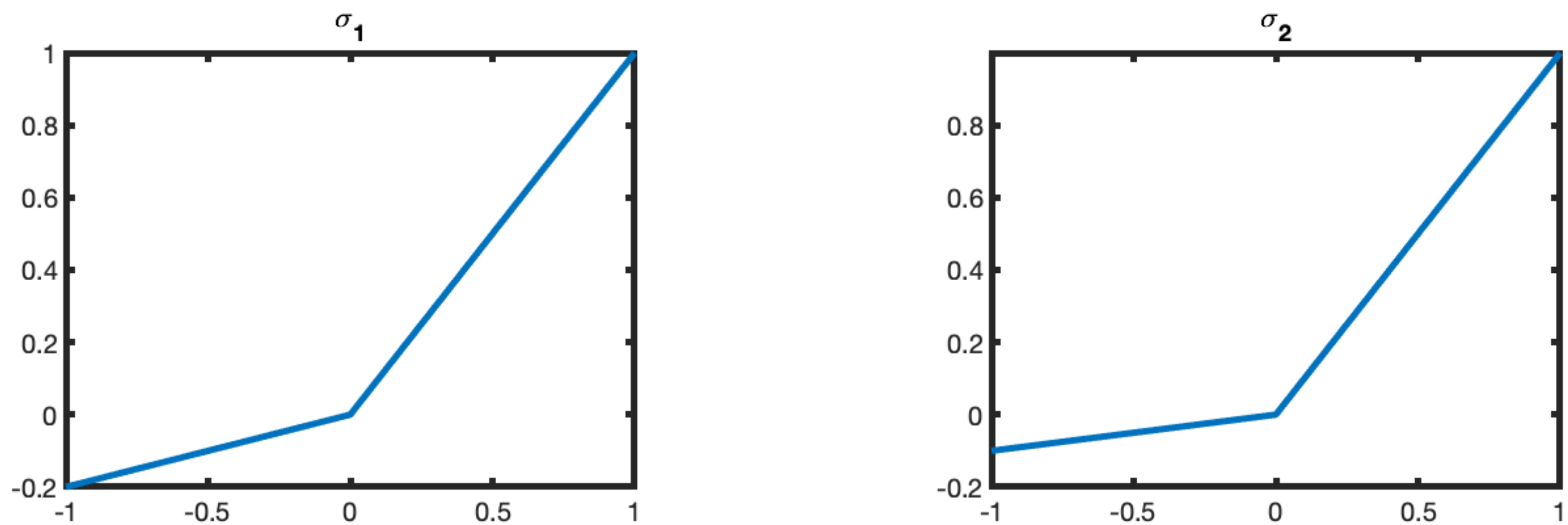
- In 1D: CPWL $\iff$ Linear spline
- Linear combination of CPWL functions $\Rightarrow$ CPWL
- Composition of two CPWL $\Rightarrow$ CPWL

$\Rightarrow$ Neural networks with linear spline activation functions are CPWL.

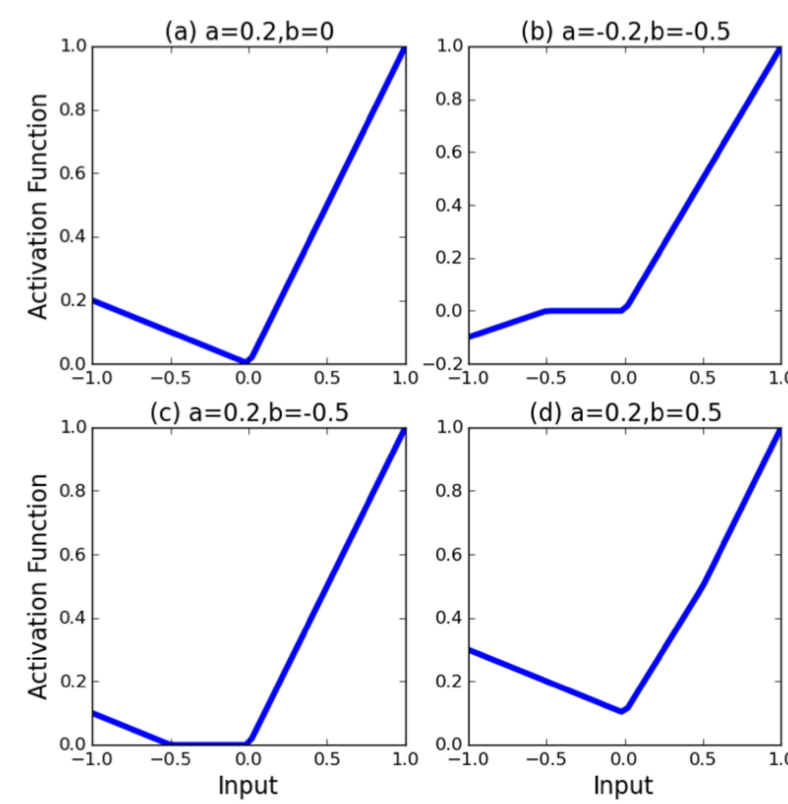
Theorem[Arora, *et al.*, 2018]: Any CPWL function $f : \mathbb{R}^d \to \mathbb{R}$ can be *exactly* represented by a ReLU neural network with at most $\lceil \log_2(d+1) \rceil + 1$ layers.

Parametric Activation Functions

- PReLU: Learn the negative slope (He *et al.* '15)



- Adaptive Piecewise Linear (APL)
  - $\sigma(x) = \mathrm{ReLU}(x) + \sum_{k=1}^{K} a_k \mathrm{ReLU}(b_k - x)$
  - $K < 10$
  - ℓ_2 regularization on a_k's and b_k's



(Agostinelli *et al.* '15)

Free-Form Activation Functions

- Deep splines: a functional framework for learning activation functions
- Principled design:
  - Preserves CPWL structure of DNNs
  - Promotes sparse activation functions
  - Controls the global Lipschitz regularity of the network
  - Efficient implementation that makes it scalable in time and memory

1D Regression with Lipschitz Regularization

- Lipschitz constant: $L(f) = \sup_{x_1 \neq x_2} \frac{|f(x_1) - f(x_2)|}{|x_1 - x_2|}$

- $\mathrm{Lip}(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} : \quad L(f) < +\infty\}$

Theorem [A. et al. '21, simplified]

There exists a linear spline solution f^* of

$$\mathcal{V}_{\mathrm{Lip}} = \arg\min_{f \in \mathrm{Lip}(\mathbb{R})} \left(\sum_{m=1}^{M} E(f(x_m), y_m) + \lambda L(f) \right)$$

with at most M knots. Moreover, we have that

$$L(f^*) = \max_{m \neq n} \left| \frac{f^*(x_m) - f^*(x_n)}{x_m - x_n} \right|.$$

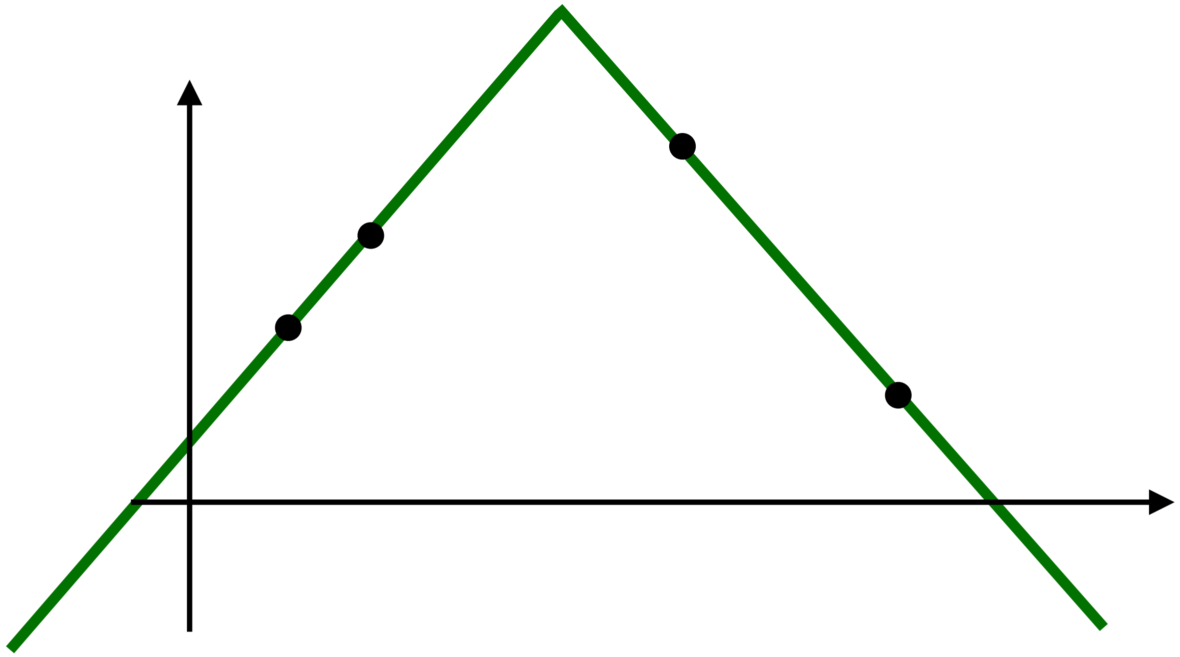
Finding The Sparsest Linear Spline Solution

- Two-stage algorithm: assume that $x_1 < \ldots < x_M$

  - Using proximal methods (*e.g. ADMM*), solve the minimization

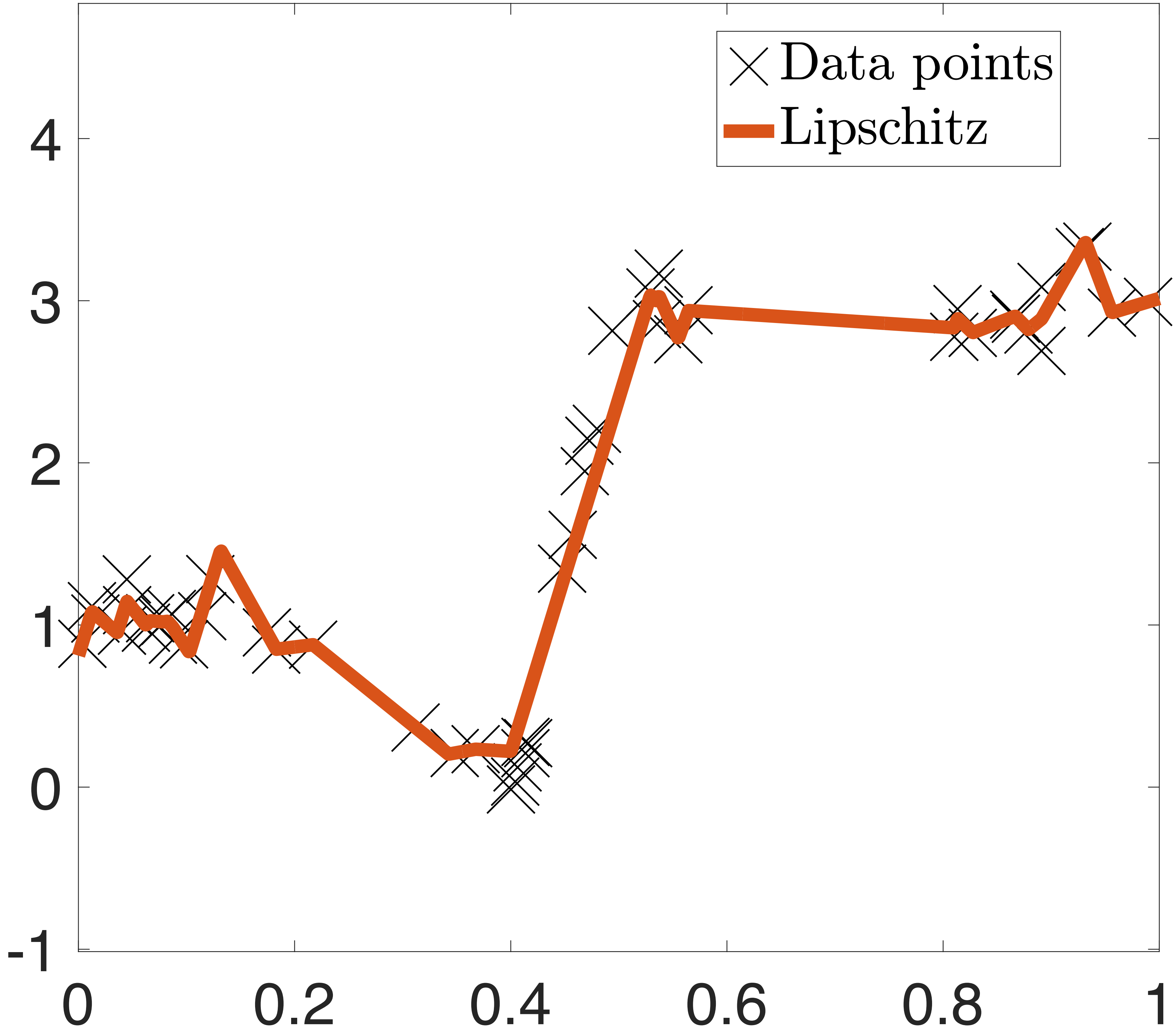
$$\arg\min_{\boldsymbol{z}\in\mathbb{R}^M} \sum_{m=1}^{M} E(y_m, z_m) + \lambda \max_{2\leq m\leq M} \left| \frac{z_m - z_{m-1}}{x_m - x_{m-1}} \right|$$

  - Find the sparsest linear spline interpolant of $(x_1, z_1^*), \ldots, (x_M, z_M^*)$.



(Debarre *et al.* '20)

Not That Sparse!



1D Regression with Sparsity

- Simple observation:

$$f(x) = ax + b + \sum_{k=1}^{K} a_k \mathrm{ReLU}(\cdot - x_k) \Rightarrow \mathrm{D}^2\{f\} = \sum_{k=1}^{K} a_k \delta(\cdot - x_k)$$

$$\Rightarrow \mathrm{TV}^{(2)}(f) = \|\mathrm{D}^2\{f\}\|_{\mathcal{M}} = \sum_{k=1}^{K} |a_k|$$

Sparsity promoting!

- Connection to Lipschitz regularity:

$$L(f) \leq \|f\|_{\mathrm{BV}^{(2)}} = \mathrm{TV}^{(2)}(f) + |f(0)| + |f(1)|$$

Theorem [Unser et al. ’17, simplified] (Debarre *et al.* '20)

There exists a linear spline solution f^* of

$$\mathcal{V}_{\mathrm{TV}^{(2)}} = \arg\min_{f \in \mathrm{BV}^{(2)}(\mathbb{R})} \left(\sum_{m=1}^{M} E(f(x_m), y_m) + \lambda \mathrm{TV}^{(2)}(f) \right)$$

with at most M knots.

Sparse + Lipschitz

- Explicit control of Lipschitz constant (Arjovsky *et al.* '17) (Bohra *et al.* '21)

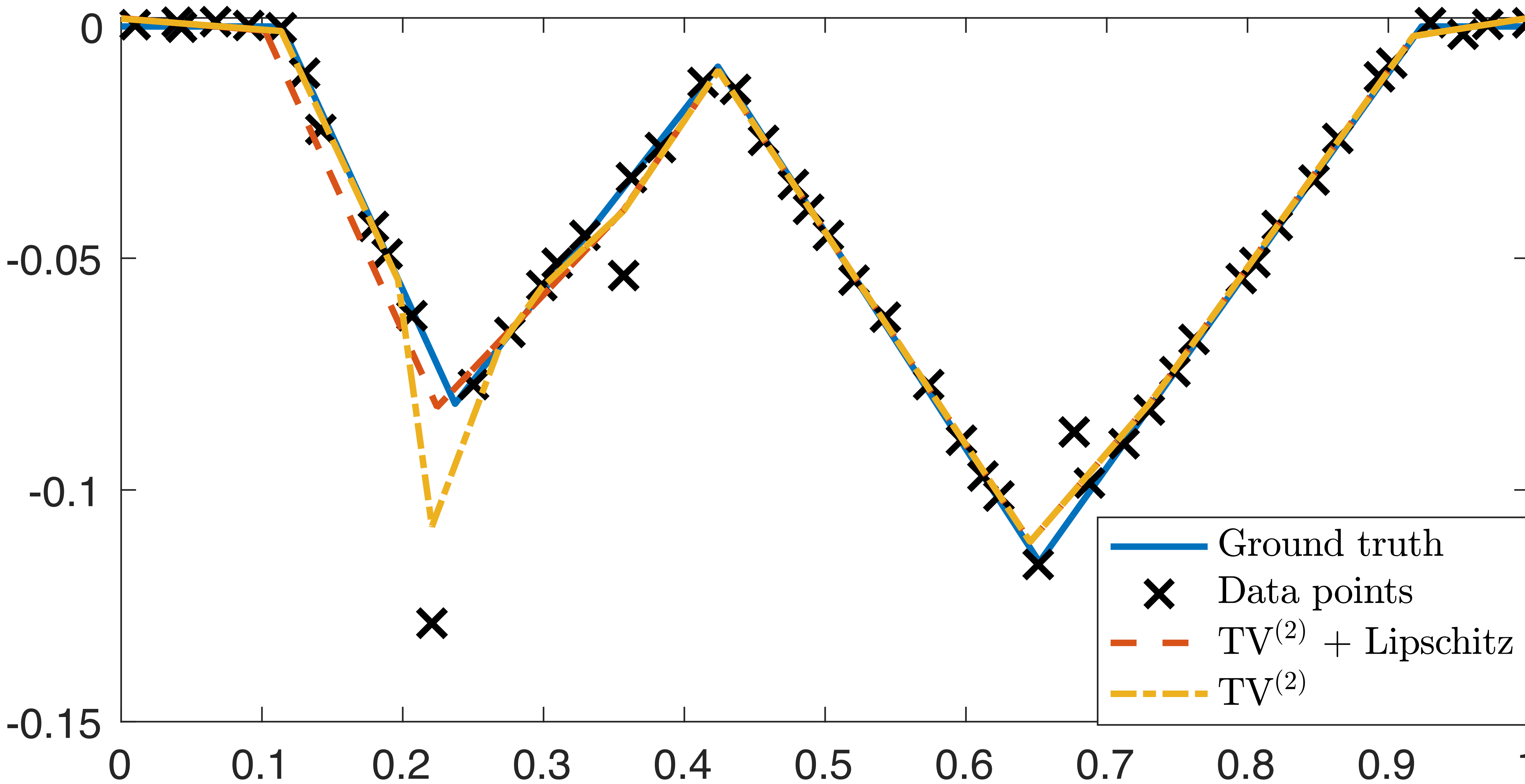
$$\mathcal{V}_{\mathrm{hyb}} = \arg\min_{f \in \mathrm{BV}^{(2)}(\mathbb{R})} \left(\sum_{m=1}^{M} \mathrm{E}(f(x_m), y_m) + \lambda \mathrm{TV}^{(2)}(f) \right), \quad \text{s.t.} \quad L(f) \leq \bar{L}$$

- $\bar{L}$: user-defined guarantee of stability

Theorem [A. et al. '21]
The solution set $\mathcal{V}_{\mathrm{hyb}}$ is a nonempty, convex and weak*-compact subset of $\mathrm{BV}^{(2)}(\mathbb{R})$ whose extreme points are linear splines with at most M knots. Moreover, there exists a unique vector $\mathbf{z}^* = (z_m)$ such that

$$\mathcal{V}_{\mathrm{hyb}} = \arg\min_{f \in \mathrm{BV}^{(2)}(\mathbb{R})} \mathrm{TV}^{(2)}(f), \quad \text{s.t.} \quad f(x_m) = z_m, 1 \leq m \leq M$$

Example



0
-0.05
-0.1
-0.15
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Ground truth
Data points
$\mathrm{TV}^{(2)}$ + Lipschitz
$\mathrm{TV}^{(2)}$

Back to DNNs

- Recall: $\mathbf{f}_{\text{deep}} = \boldsymbol{\sigma}_L \circ \mathbf{W}_L \circ \cdots \circ \boldsymbol{\sigma}_1 \circ \mathbf{W}_1 : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$

- $\boldsymbol{\sigma} = (\sigma_n) \in \mathrm{BV}^{(2)}(\mathbb{R})^N \Rightarrow \|\boldsymbol{\sigma}\|_{\mathrm{BV}^{(2)}} = \sum_{n=1}^{N} \|\sigma_n\|_{\mathrm{BV}^{(2)}}$

Theorem [A. et al. '20]
Any feed-forward fully-connected deep neural network with second-order bounded activation functions is Lipschitz continuous. Moreover, the Lipschitz constant of $\mathbf{f}_{\text{deep}} : (\mathbb{R}^{N_0}, \|\cdot\|_2) \to (\mathbb{R}^{N_L}, \|\cdot\|_2)$ is upper-bounded by

$$L(\mathbf{f}_{\text{deep}}) \leq \left(\prod_{l=1}^{L} \|\mathbf{W}_l\|_F\right) \cdot \left(\prod_{l=1}^{L} \|\boldsymbol{\sigma}_l\|_{\mathrm{BV}^{(2)}}\right)$$

Deep Splines

Theorem [A. et al. '20] (Unser '19)

There exists an optimal configuration that minimizes the cost functional

$$\mathcal{J}(\mathbf{f}_{\mathrm{deep}}) = \sum_{m=1}^{M} E\big(\boldsymbol{y}_m, \mathbf{f}_{\mathrm{deep}}(\boldsymbol{x}_m)\big) + \sum_{l=1}^{L} \mu_l \|\mathbf{W}_l\|_F^2 + \sum_{l=1}^{L} \lambda_l \|\boldsymbol{\sigma}_l\|_{\mathrm{BV}^{(2)}}$$

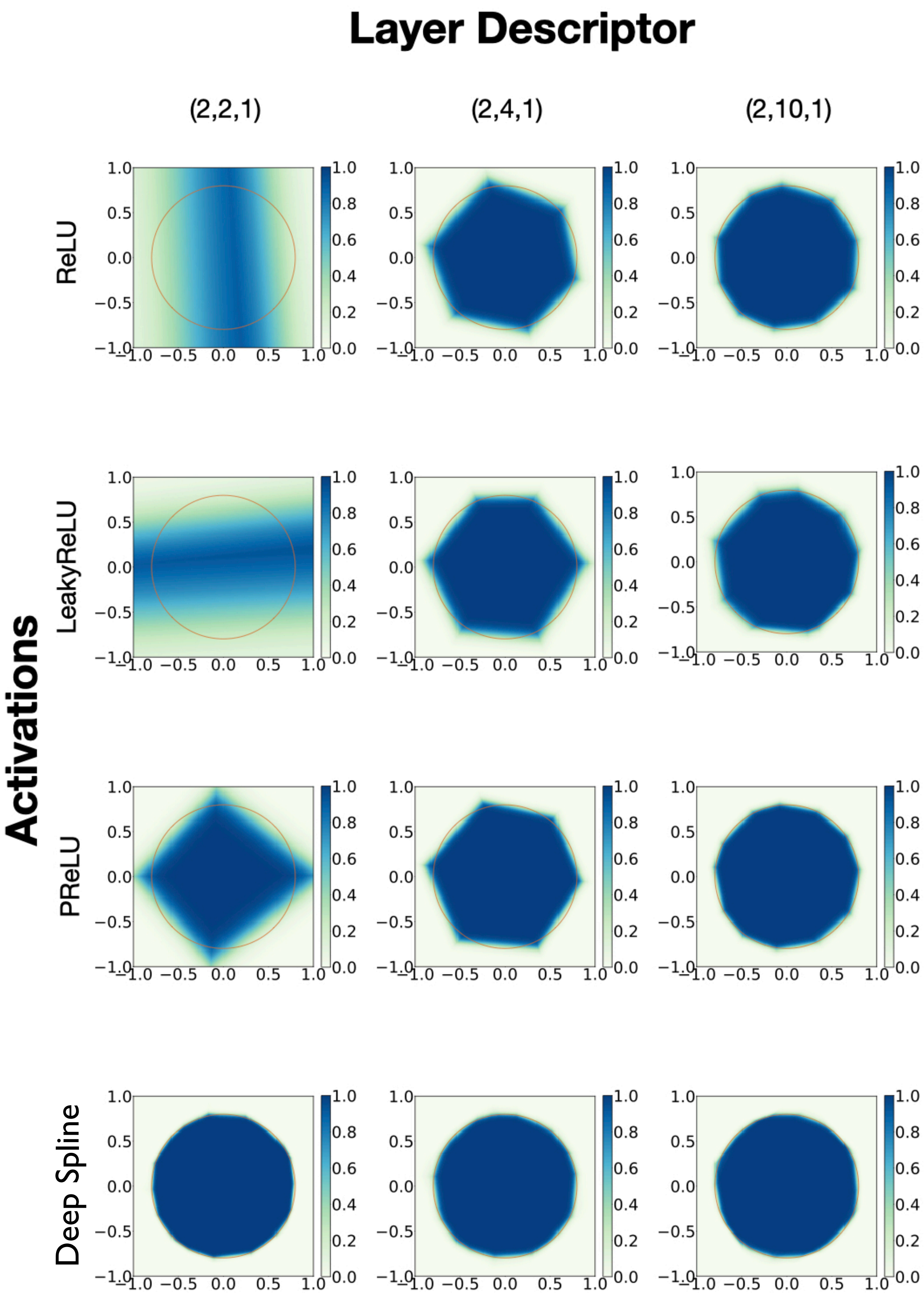
whose activation functions are linear splines with at most M knots.

Moreover, any local minima of the above problem satisfies

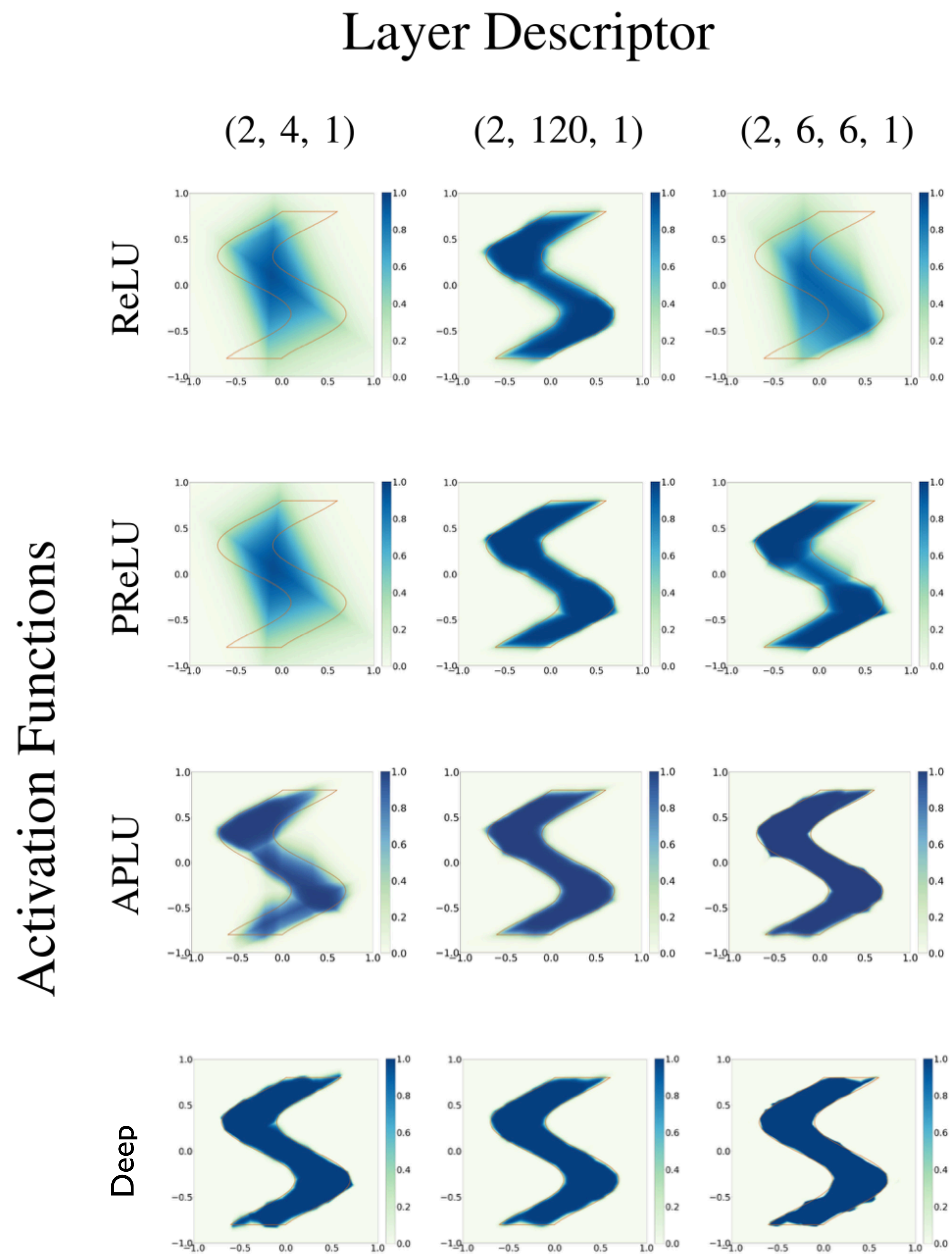
$$\lambda_l \|\boldsymbol{\sigma}_l\|_{\mathrm{BV}^{(2)}} = 2\mu_{l+1}\|\mathbf{W}_{l+1}\|_F^2, \quad l = 1, \dots, L-1.$$

- Open-source software: github.com/joaquimcampos/DeepSplines

Examples



Layer Descriptor
(2,2,1)
(2,4,1)
(2,10,1)
Activations
ReLU
LeakyReLU
PReLU
Deep Spline



Layer Descriptor
(2, 4, 1)
(2, 120, 1)
(2, 6, 6, 1)
Activation Functions
ReLU
PReLU
APLU
Deep

Examples

TABLE 2 NIN Error Rates on CIFAR-10 and CIFAR-100

Activation function	CIFAR-10	CIFAR-100
ReLU	8.78%	32.44%
APLU	8.71%	31.74%
B-spline	8.29%	30.43%

TABLE 3 ResNet Error Rates on CIFAR-10 and CIFAR-100

Activation function	CIFAR-10	CIFAR-100
ReLU	6.31%	29.02%
APLU	6.45%	28.85%
B-spline	6.02%	28.24%

TABLE 4 B-Splines *vs.* Gridded ReLUs *vs.* APLUs

Architecture, Nb. coefficients	Memory (megabytes)	Time per epoch (seconds)
B-splines, $K = 9$	1132	44.92
B-splines, $K = 29$	1133	41.89
B-splines, $K = 499$	1299	41.19
Gridded ReLUs, $K = 9$	3313	49.86
Gridded ReLUs, $K = 29$	9616	81.21
APLUs, $K = 9$	3316	49.72
APLUs, $K = 29$	9618	87.34

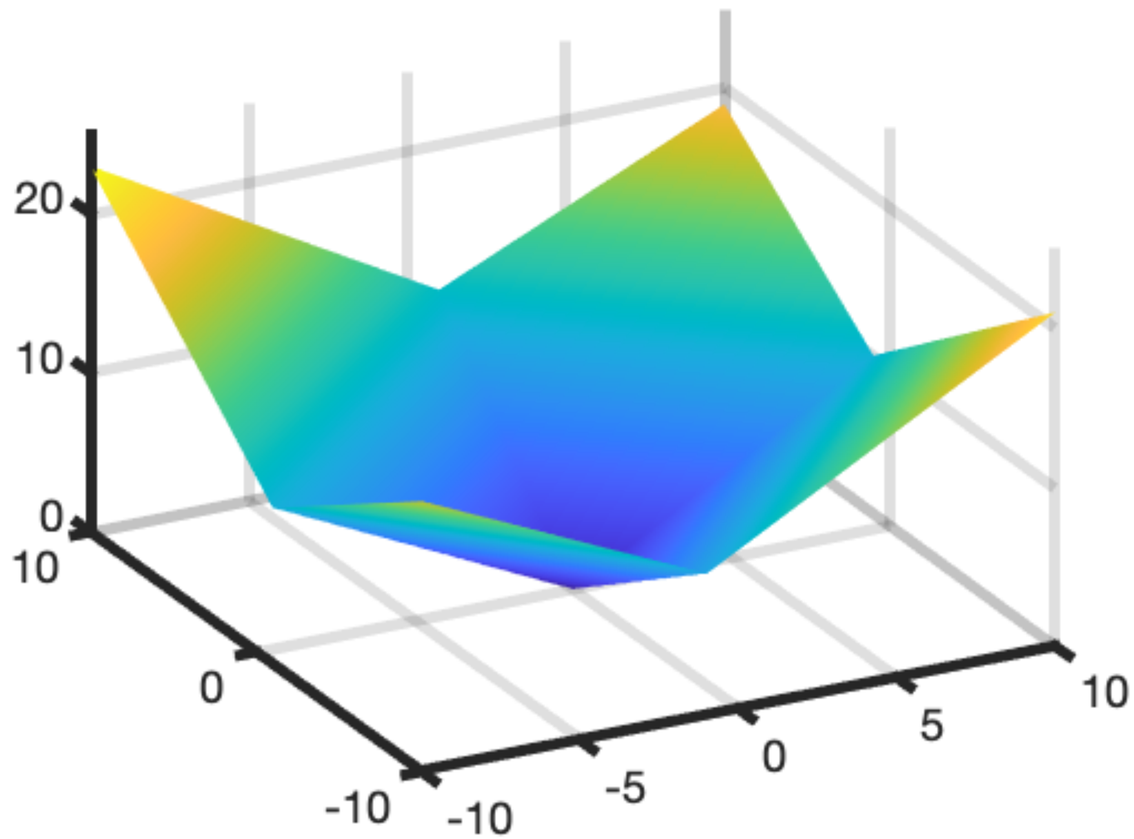
For the gridded ReLU and APLU networks, the maximum number of knots allowed by the GPU memory is 31.

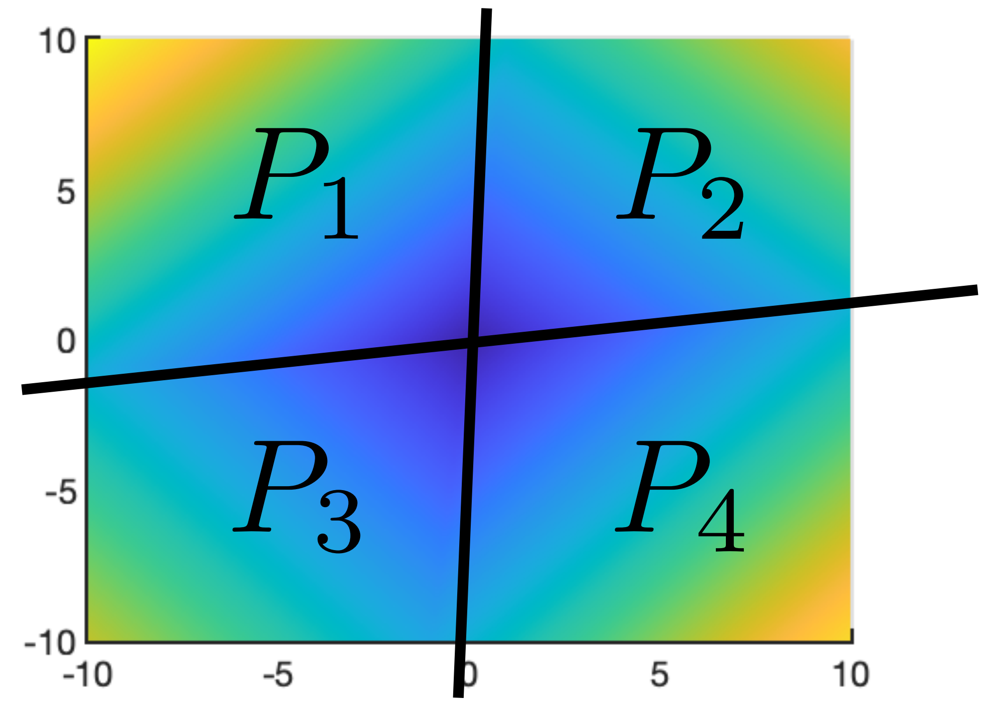
Source: P. Bohra, J. Campos, H. Gupta, S. Aziznejad, M. Unser, "**Learning Activation Functions in Deep (Spline) Neural Networks**," IEEE Open Journal of Signal Processing, 2020.

OUTLINE

- **Introduction** ✔
- **Learning over Banach spaces** ✔
  - Theory of Banach spaces
  - General representer theorem
  - Application: Sparse multikernel regression
- **Learning activation functions of DNNs** ✔
  - One-dimensional learning
  - Deep splines
- **Going to higher dimensions**
  - Hessian-based regularization
- **Future works**

CPWL Functions Revisited





- Hessian of CPWL functions has Hausdorff dimension $= (d-1)$
- Intuition: Schatten-1 norm regularization promotes low-rank matrices

Related Literature

- R. Parhi, R.D. Nowak, "Banach space representer theorems for neural networks and ridge splines," *Journal of Machine Learning Research*, 2021.
- R. Parhi, R.D. Nowak, "What Kinds of Functions do Deep Neural Networks Learn? Insights from Variational Spline Theory," *ArXiv*, 2021.
- P. Savarese, I. Evron, D. Soudry, N. Srebro, "How do infinite width bounded norm networks look in function space?," *Conference on Learning Theory*, 2019.
- G. Ongie, R. Willett, D. Soudry, N. Srebro, "A function space view of bounded norm infinite width ReLU nets: The multivariate case," *ArXiv,* 2019.

Hessian-Schatten Total Variation

- Informal definition $$\mathrm{HTV}_p(f) = \int_{\mathbb{R}^d} \|\mathrm{H}\{f\}(\boldsymbol{x})\|_{S_p} \mathrm{d}\boldsymbol{x}$$

- Hessian of CPWL functions is not defined pointwise!

Definition [A. et al. '21]

Let $p \in [1, +\infty]$ and $q = p/(p-1)$. The Hessian-Schatten total-variation (HTV) of any $f : \mathbb{R}^d \to \mathbb{R}$

$$\mathrm{HTV}_p(f) = \sup \left\{ \langle \mathrm{H}\{f\}, \mathbf{F} \rangle : \mathbf{F} = [f_{i,j}], f_{i,j} \in \mathcal{C}_0(\mathbb{R}^d), \|\mathbf{F}(\boldsymbol{x})\|_{S_q} \leq 1 \forall \boldsymbol{x} \in \mathbb{R}^d \right\}.$$

Hessian-Schatten Total-Variation

Theorem [A. et al. '21]

1. If $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable, then

$$\mathrm{HTV}_p(f) = \int_{\mathbb{R}^d} \|\mathrm{H}\{f\}(\boldsymbol{x})\|_{S_p} \mathrm{d}\boldsymbol{x}.$$

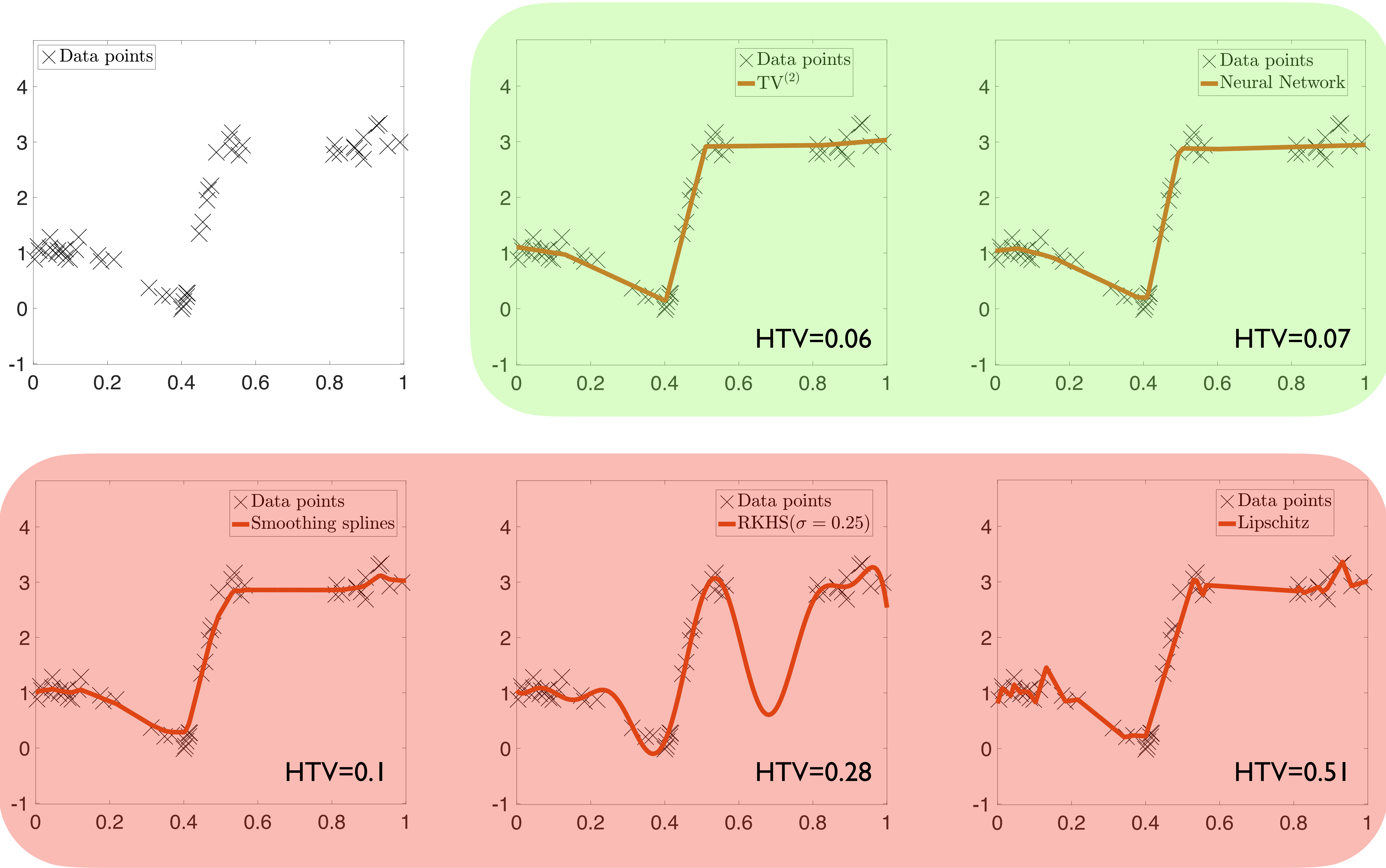
2. Let f be a CPWL function with linear regions $P_1, \ldots, P_N$ so that $\nabla f\big|_{P_n} = \boldsymbol{a}_n \in \mathbb{R}^d$ for $n = 1, \ldots, N$. Then

$$\mathrm{HTV}_p(f) = \sum_{m<n} \|\boldsymbol{a}_n - \boldsymbol{a}_m\|_2 \mathcal{H}^{d-1}(P_n \cap P_m),$$

where $\mathcal{H}^{d-1}$ denotes the $(d-1)$-dimensional Hausdorff measure.

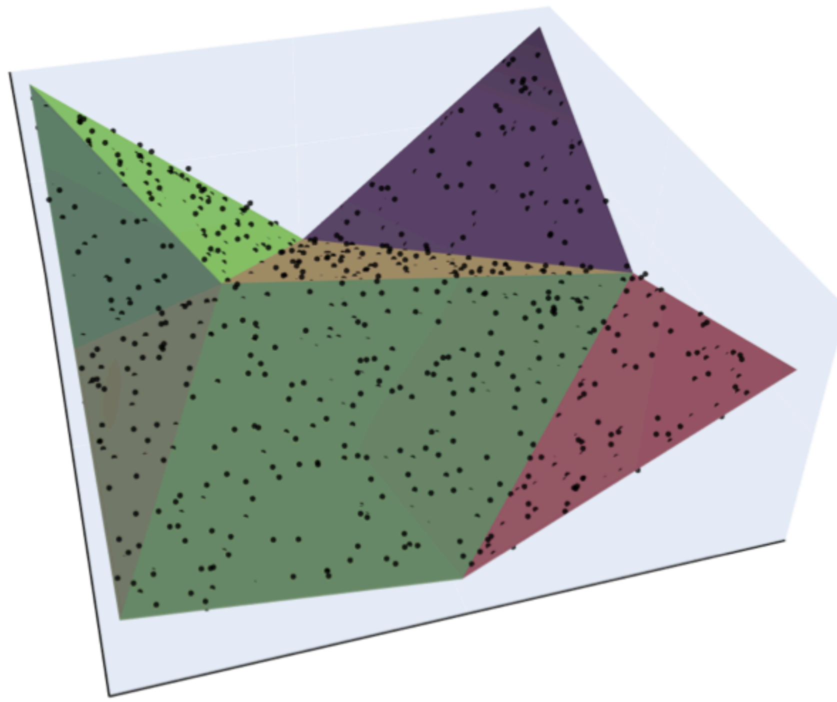
- Proof of 1: Duality mapping of Schatten norms (A.-Unser '21)

Example: HTV As a Complexity Measure

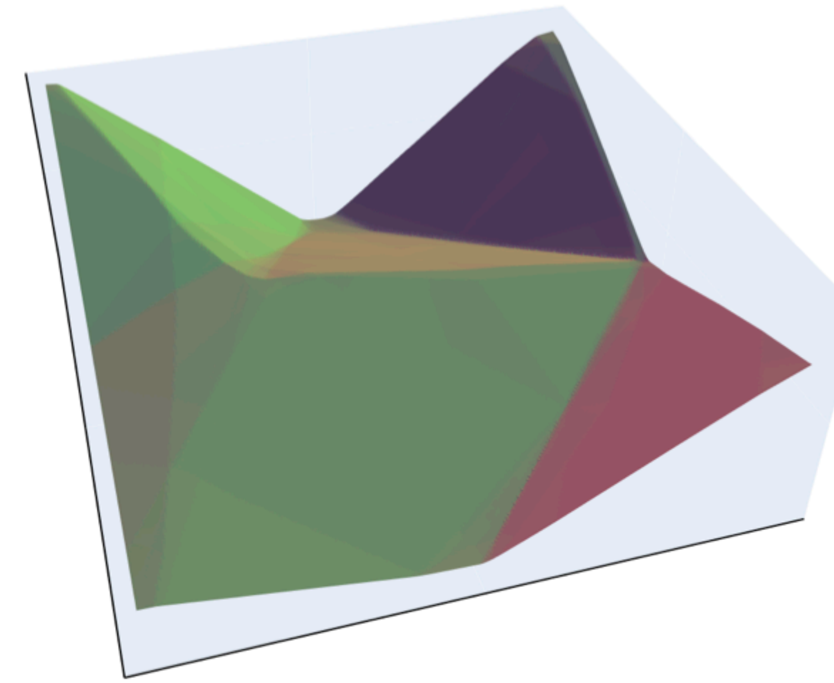


Example: HTV As a Complexity Measure

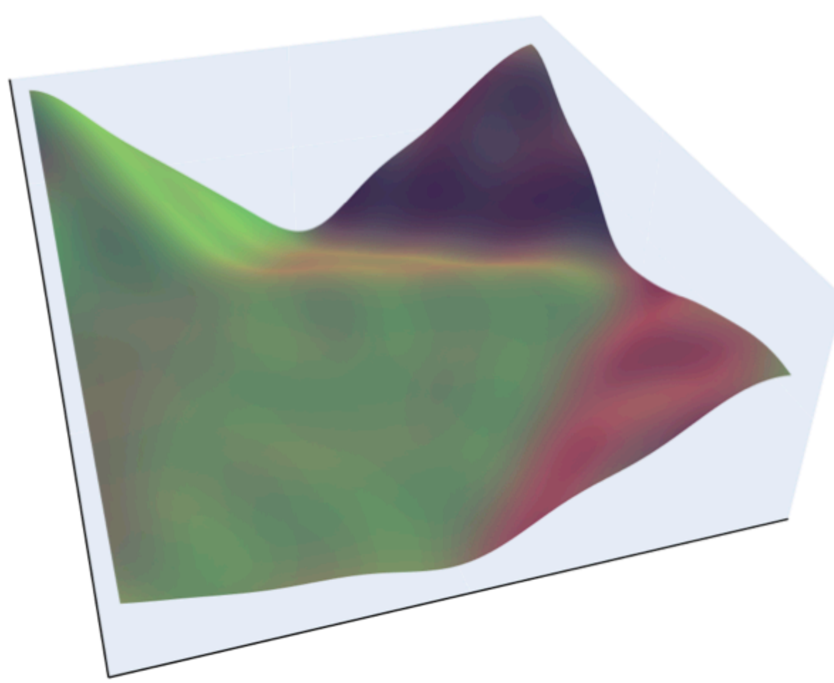
Target function
HTV = 6.98
+
Noisy training data



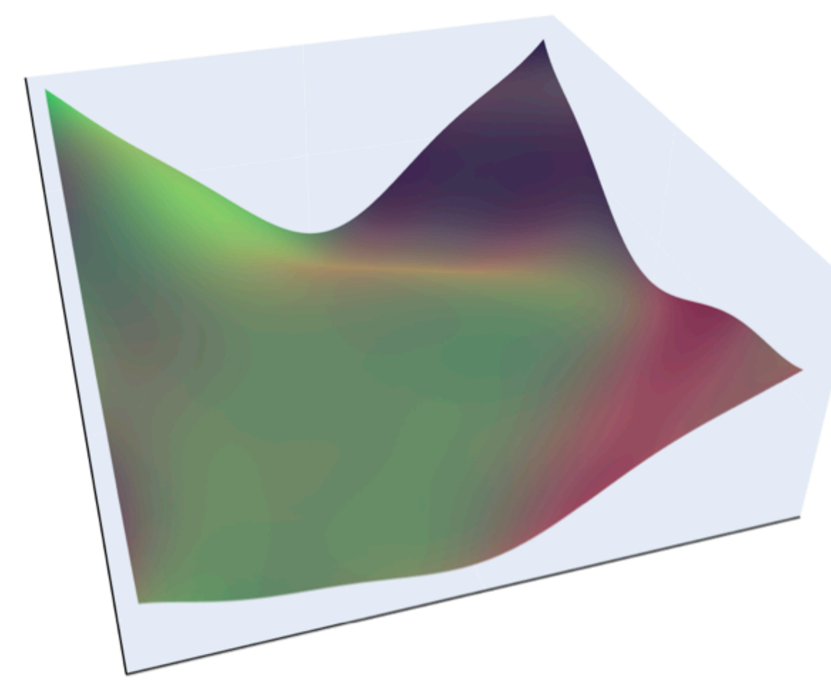
ReLU neural network
(2,40,40,40,40,1)
Weight decay= 5e-5
MSE= 2.36e-5
HTV= 8.1



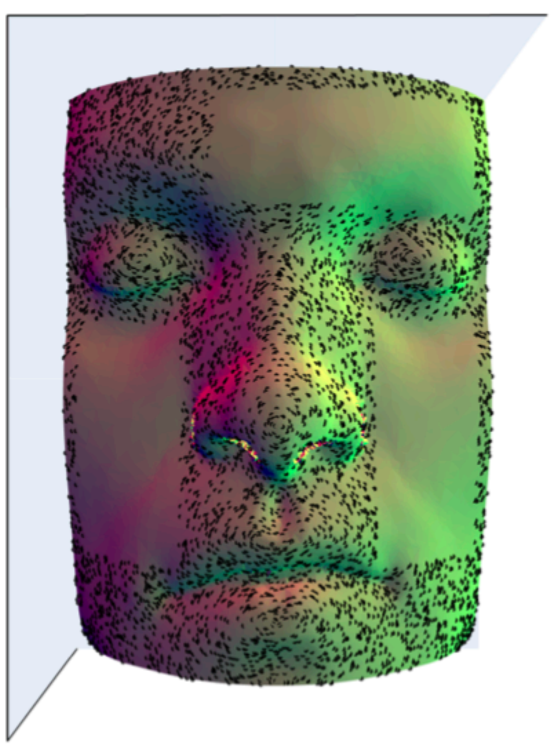
Gaussian RBF
Sigma= 0.41
Lambda= 5e-6
MSE= 6.58e-5
HTV_{10}= 10.44



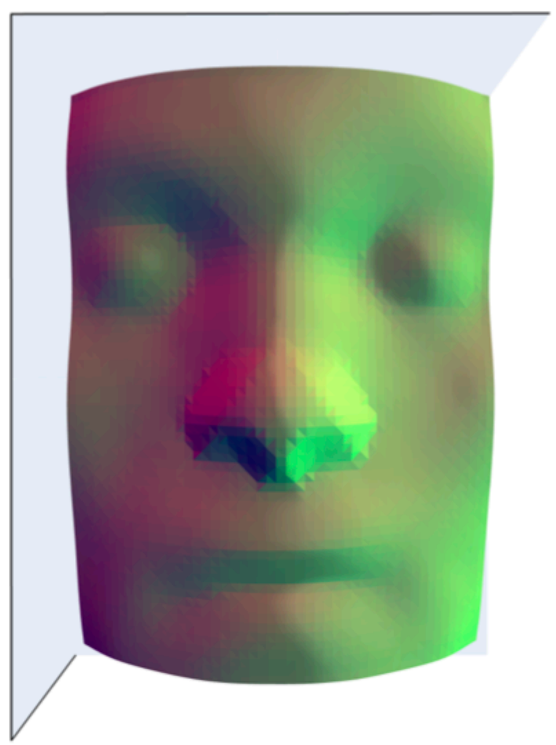
Gaussian RBF
Sigma= 0.71
Lambda= 1e-2
MSE= 1.69 e-4
HTV_{10}= 8.2



Example: HTV As a Complexity Measure

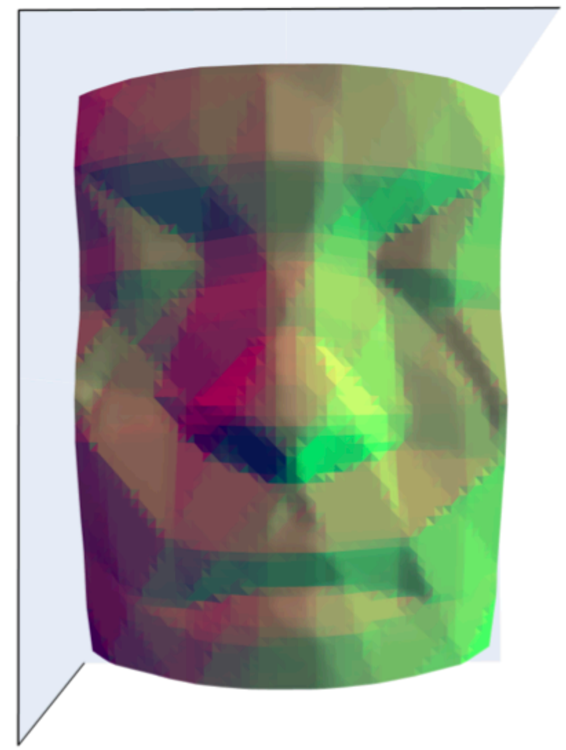


Target function
+
M=5000 training data

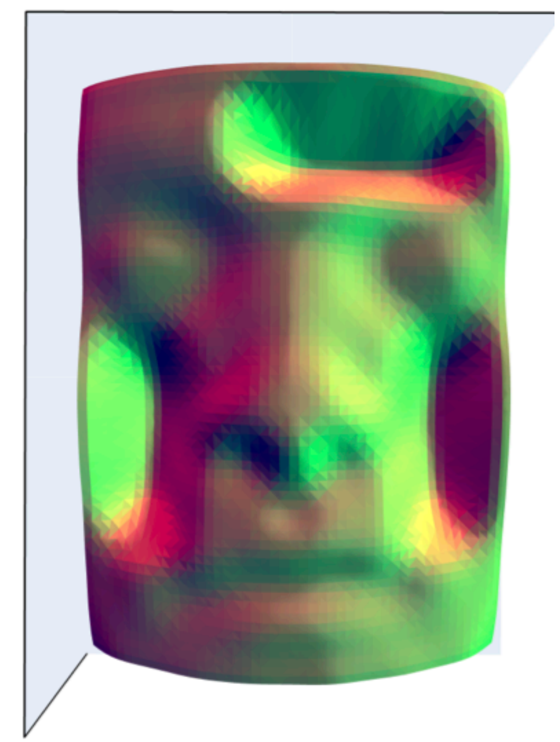


HTV Min

Train SNR = 39.4 dB
Test SNR = 34.84 dB
HTV = 8.9



ReLU neural network
(2,40,40,40,40,1)
Train SNR = 39.6 dB
Test SNR = 33.0 dB
HTV= 10.8



Gaussian RBF
Sigma= 0.16
Train SNR = 39.4 dB
Test SNR = 13.6 dB
HTV_1= 24.3

Source: J. Campos, S. Aziznejad, M. Unser, "**Learning of Continuous and Piecewise-Linear Functions with Hessian Total-Variation Regularization**," submitted, 2021.

Conclusion

- A general framework for learning over Banach spaces
  - Application: Sparse multikernel regression
- Learning sparse and Lipschitz-regular 1D mappings
  - Application: Deep splines
- Learning CPWL functions in higher dimensions
  - Defining a Hessian-based regularization functional

Selected references

- Optimization Over Banach Spaces
  - M. Unser, "A Unifying Representer Theorem for Inverse Problems and Machine Learning," *FoCM,* 2021.
  - M. Unser, **S. Aziznejad**, "Convex Optimization in Sums of Banach Spaces," *ACHA,* 2022.
  - **S. Aziznejad**, M. Unser, "Multikernel Regression with Sparsity Constraint," SIMODS, 2021.
- Deep Splines
  - M. Unser, ``A Representer Theorem for Deep Neural Networks", *JMLR*, 2019.
  - **S. Aziznejad**, H. Gupta, J. Campos, M. Unser, "Deep Neural Networks with Trainable Activations and Controlled Lipschitz Constant," *IEEE TSP,* 2020.
  - P. Bohra, J. Campos, H. Gupta, **S. Aziznejad**, M. Unser, "Learning Activation Functions in Deep (Spline) Neural Networks," *IEEE OJSP,* 2020.
  - **S. Aziznejad**, T. Debarre, M. Unser, "Robust and Sparse Regression Models for One-dimensional Data", submitted, 2021.
  - T. Debarre, Q. Denoyelle, J. Fageot, M. Unser, "Sparsest Continuous and Piecewise Representation of Data", *submitted,* 2020.
- Hessian-Based Regularization
  - **S. Aziznejad**, M. Unser, "Duality Mapping for Schatten Matrix Norms", *Numerical Functional Analysis and Optimization,* 2021.
  - **S. Aziznejad**, M. Unser, "Measuring Complexity of Learning Schemes with Hessian-Schatten Total Variation", *submitted,* 2021.
  - J. Campos, **S. Aziznejad**, M. Unser, "Learning Continuous and Piecewise Linear Functions with Hessian-Schatten Total-Variation Regularization", *submitted,* 2021.