Optimization Over Banach Spaces: A Unified View on Supervised Learning and Inverse Problems

Shayan Aziznejad

Biomedical Imaging Group
EPFL, Lausanne, Switzerland

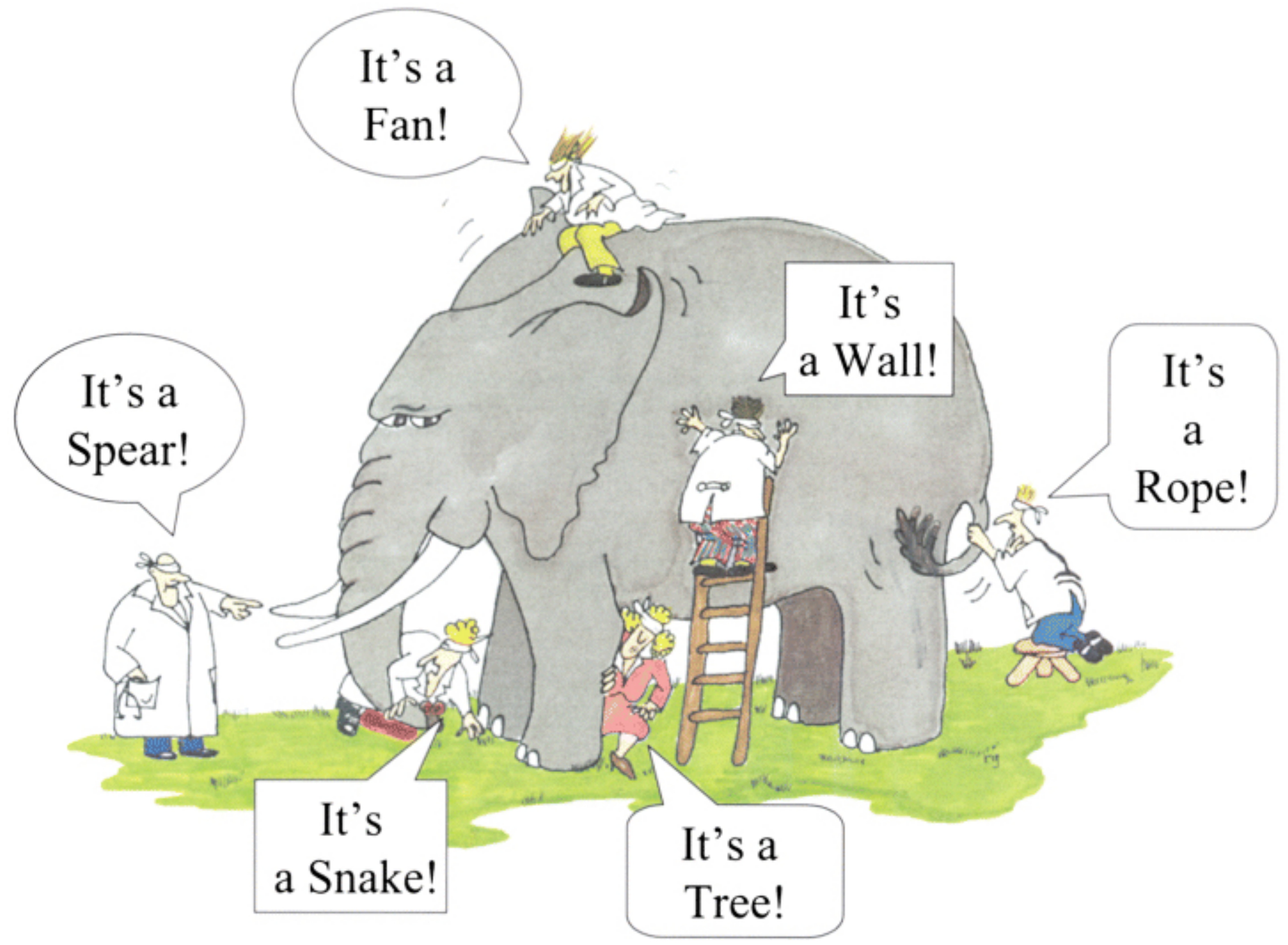
Jury Members:

- Prof. D. Van De Ville, president
- Prof. M. Unser, thesis director
- Prof. A. C. Hansen, external examiner
- Prof. G. Peyré, external examiner
- Prof. V. Panaretos, internal examiner

PhD defense
May 2, 2022

Inverse Problems

- Recovering an unknown signal from a collection of observations



Blind men and an elephant

- The mathematical setting of interest

  - Continuous-domain problems

    $f : \mathbb{R}^d \to \mathbb{R}$: Signal of interest

    $f \in \mathcal{F}(\mathbb{R}^d)$: Infinite-dimensional search space

  - Finitely many noisy observations

    $\boldsymbol{y} = (y_1, \dots, y_M) \in \mathbb{R}^M$: Measurement vector

    $y_m \approx \nu_m(f), \quad m = 1, \dots, M$: Forward model

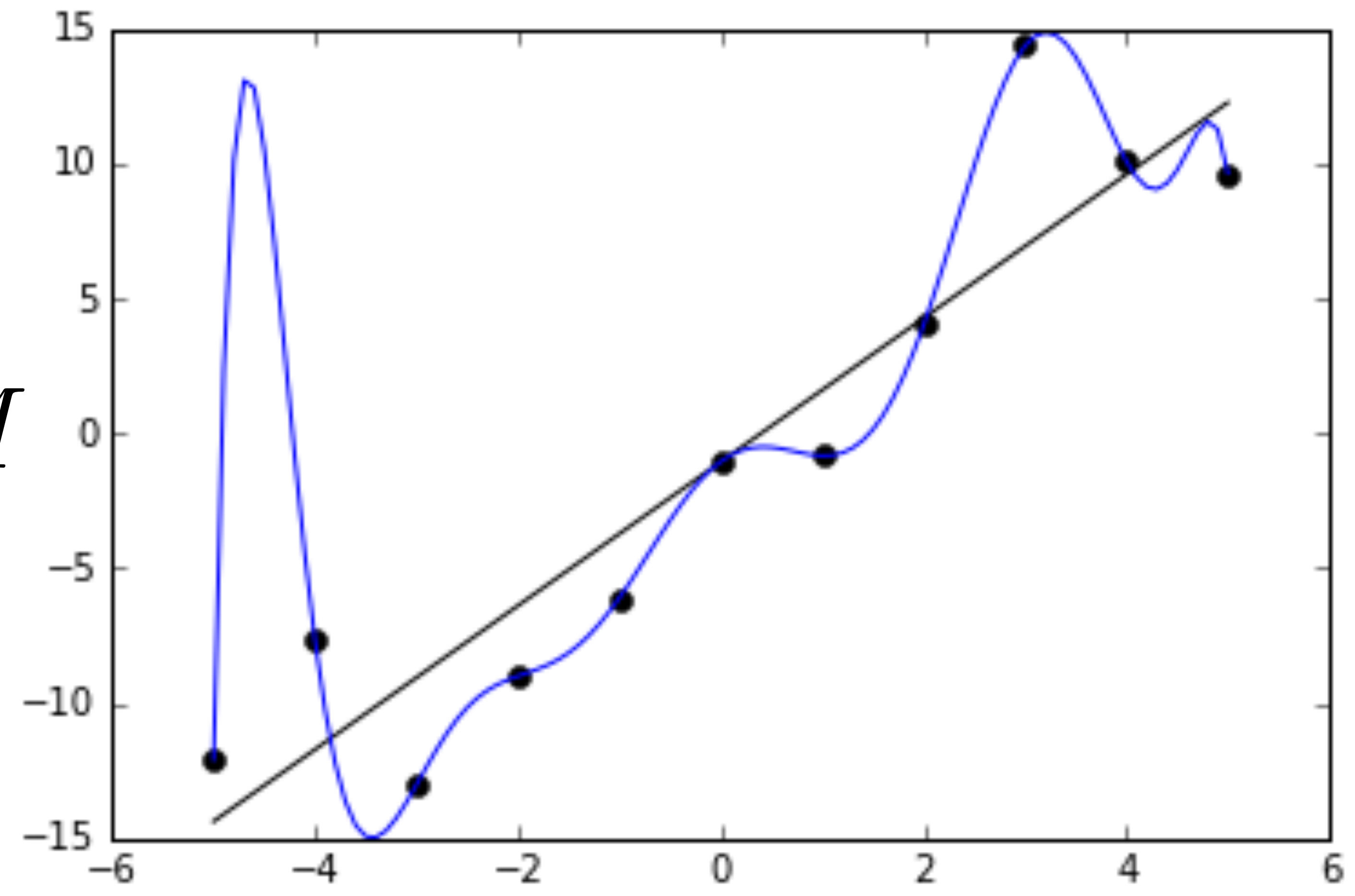
  - Linear forward model

    $\boldsymbol{\nu} = (\nu_m) : \mathcal{F}(\mathbb{R}^d) \to \mathbb{R}^M$: Continuous vector-valued linear functional

Supervised Learning

Without Overfitting!

- Training data: $\{(x_m, y_m)\}_{m=1}^{M} \subseteq \mathcal{X} \times \mathcal{Y}$



Source: en.wikipedia.org/wiki/Overfitting

- Goal: Find $f : \mathcal{X} \to \mathcal{Y}$ such that $f(x_m) \approx y_m$ for $m = 1, \dots, M$

- Nonparametric regression
  - $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$
  - $f \in \mathcal{F}(\mathbb{R}^d)$

- Supervised learning as a special linear inverse problem
  - $\boldsymbol{\nu} : f \mapsto (f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_M)) \in \mathbb{R}^M$

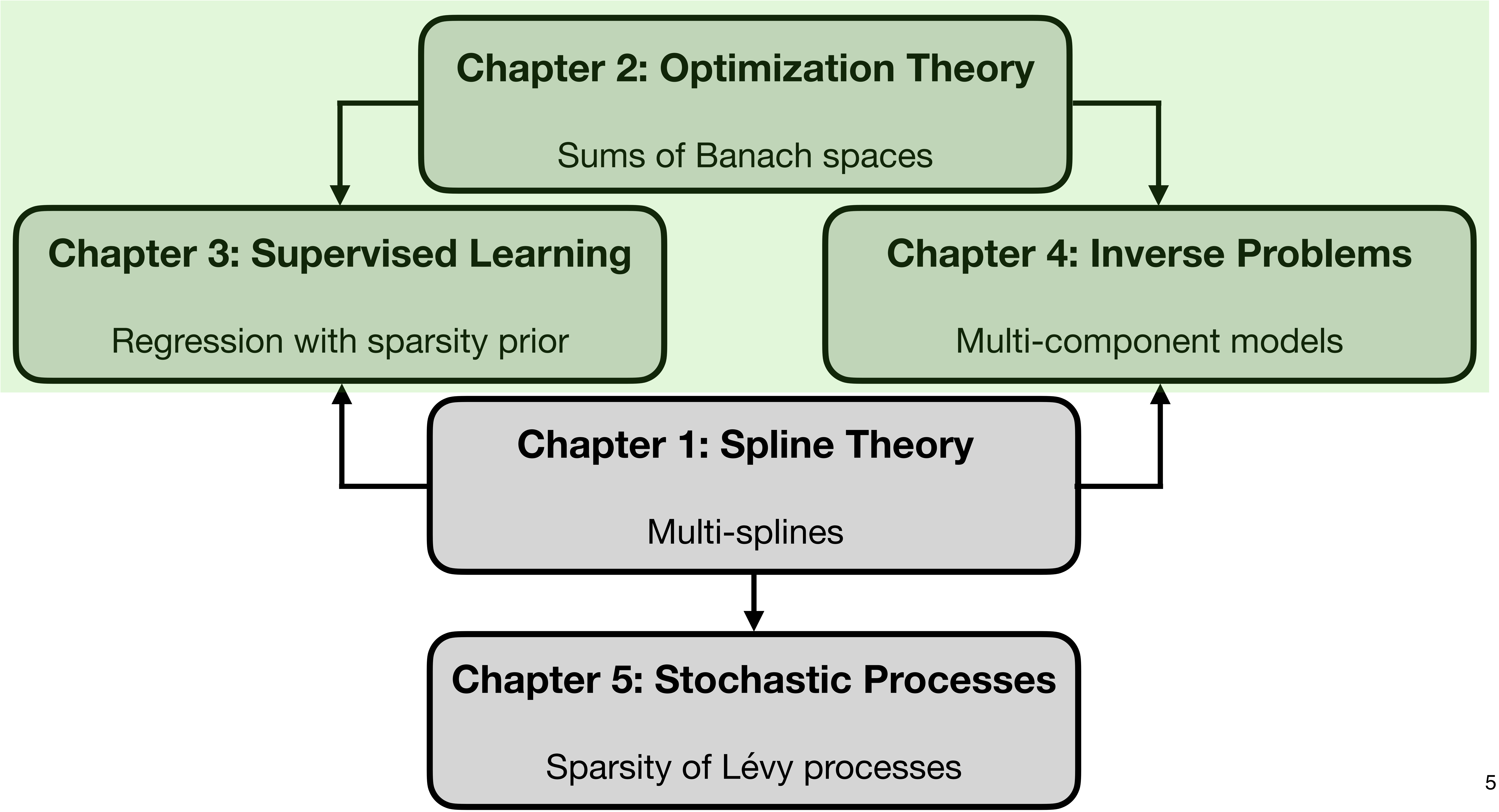
$\nu_m = \delta_{\boldsymbol{x}_m} : \mathcal{F}(\mathbb{R}^d) \to \mathbb{R} : f \mapsto f(\boldsymbol{x}_m)$: Sampling functional

Variational Formulation of Inverse Problems

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

- $E : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$: Convex loss function
  - Penalizes the data discrepancy
  - Related to the noise model
  - *e.g.* Quadratic loss $E(y, z) = (y - z)^2$
- $\mathcal{R} : \mathcal{F}(\mathbb{R}^d) \to \mathbb{R}_{\geq 0}$: Regularization functional
  - Enforces prior knowledge on the reconstructed signal
  - Related to the signal model
  - *e.g.* Tikhonov, total-variation (TV)
- $\mathcal{F}(\mathbb{R}^d)$: Hilbert space ✔
- $\mathcal{F}(\mathbb{R}^d)$: Banach space?

Outline of the Thesis



Part I: Optimization over Banach Spaces

$$\mathcal{V} = \arg\min_{f \in \mathcal{F}} \|\boldsymbol{\nu}(f) - \boldsymbol{y}\|_2^2 + \lambda \mathcal{R}(f)$$

- General representer theorem [Unser'21]:
  - Full characterization when $\mathcal{F} = \mathcal{X}'$ and $\mathcal{R}(f) = \|f\|_{\mathcal{X}'}$
  - $\mathrm{Ext}(\mathcal{V})$: Linear combination of at most M extreme points of $B_{\mathcal{X}'}$
- Characterizing the solution set $\mathcal{V}$ in two different scenarios
  1. Direct-product structure: $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$, $\mathcal{F} = \mathcal{X}'$ and $\mathcal{R}(f) = \|f\|_{\mathcal{X}'}$
  2. Minimization of seminorms: $\mathcal{F} = \mathcal{U}' \oplus \mathcal{N}'$ and $\mathcal{R}(f) = \|\mathrm{Proj}_{\mathcal{U}'}(f)\|_{\mathcal{U}'}$
- Relevant publication

- M. Unser, **S. Aziznejad**, "Convex optimization in sums of Banach spaces," *Applied and Computational Harmonic Analysis*, *2022*.

Optimization over Direct-Product Spaces

Theorem [Unser-A.'22, simplified]

- $(\mathcal{X}_n, \|\cdot\|_{\mathcal{X}_n}), n = 1, \dots, N$: Banach spaces
- $(\mathcal{X}, \|\cdot\|_{\mathcal{X}}) = (\mathcal{X}_1 \times \cdots \times \mathcal{X}_N)_\infty$: Direct-product search space

$$\|(f_1, \dots, f_N)\|_{\mathcal{X}} = \max(\|f_1\|_{\mathcal{X}_1}, \dots, \|f_N\|_{\mathcal{X}_N})$$

- $\boldsymbol{\nu} = (\nu_m) : \mathcal{X}' \to \mathbb{R}^M$: Weak*-continuous

Then, the solution set

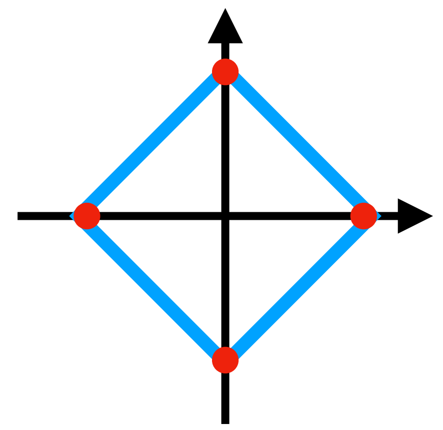
$$\mathcal{V} = \arg\min_{f \in \mathcal{X}'} \|\boldsymbol{\nu}(f) - \boldsymbol{y}\|_2^2 + \lambda \|f\|_{\mathcal{X}'}$$

is nonempty, convex and weak*-compact. Moreover

1. $\mathrm{Ext}(\mathcal{V}|_{\mathcal{X}'_n})$: linear combination of K_n extreme points of $B_{\mathcal{X}'_n}$
2. $\sum_{n=1}^N K_n \leq M$.

Sketch of proof

1. Topological structure of the search space
   - $\mathcal{X}' = \mathcal{X}'_1 \times \cdots \times \mathcal{X}'_N$
   - $\|(f_n)\|_{\mathcal{X}'} = \sum_{n=1}^N \|f_n\|_{\mathcal{X}'_n}$
2. Topological structure of $\mathcal{V}$
   - General representer theorem [Unser'21]
3. $e = (e_n) \in \mathrm{Ext}(B_{\mathcal{X}'})$ if and only if
   - $e_n \in \mathrm{Ext}(B_{\mathcal{X}'_n})$ for $n = 1, \dots, N$
   - $(\|e_1\|_{\mathcal{X}'_1}, \dots, \|e_N\|_{\mathcal{X}'_N}) \in \mathrm{Ext}(B_1)$
4. Extreme points of the unit ℓ_1 ball in $\mathbb{R}^N$
   - $\pm \mathbf{e}_n = (0, \dots, \pm 1, \dots, 0) \subseteq \mathbb{R}^N$



Part I: Optimization over Banach Spaces

$$\mathcal{V} = \arg\min_{f \in \mathcal{F}} \|\boldsymbol{\nu}(f) - \boldsymbol{y}\|_2^2 + \lambda \mathcal{R}(f)$$

- General representer theorem [Unser'21]:
  - Full characterization when $\mathcal{F} = \mathcal{X}'$ and $\mathcal{R}(f) = \|f\|_{\mathcal{X}'}$
  - $\mathrm{Ext}(\mathcal{V})$: Linear combination of at most M extreme points of $B_{\mathcal{X}'}$
- Characterizing the solution set $\mathcal{V}$ in two different scenarios
  1. Direct-product structure: $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$, $\mathcal{F} = \mathcal{X}'$ and $\mathcal{R}(f) = \|f\|_{\mathcal{X}'}$
  2. Minimization of seminorms: $\mathcal{F} = \mathcal{U}' \oplus \mathcal{N}'$ and $\mathcal{R}(f) = \|\mathrm{Proj}_{\mathcal{U}'}(f)\|_{\mathcal{U}'}$
- Relevant publication

- M. Unser, **S. Aziznejad**, "Convex optimization in sums of Banach spaces," *Applied and Computational Harmonic Analysis*, *2022*.

Minimization of Seminorms

Theorem [Unser-A.'22]

- $\mathcal{X} = \mathcal{U} \oplus \mathcal{N}$ with $\dim(\mathcal{N}) = N_0 < +\infty$
- $\boldsymbol{\nu} = (\nu_m) : \mathcal{X}' \to \mathbb{R}^M$: invertible over $\mathcal{N}'$

Then, the solution set

$$\mathcal{V} = \arg\min_{f \in \mathcal{X}'} \|\boldsymbol{\nu}(f) - \boldsymbol{y}\|_2^2 + \lambda \|\mathrm{Proj}_{\mathcal{U}'}(f)\|_{\mathcal{U}'}$$

is nonempty, convex and weak*-compact.

Moreover for any $f \in \mathrm{Ext}(\mathcal{V})$, we have that

$$f = \sum_{k=1}^{K_0} c_k e_k + p,$$

where $K_0 \leq (M - N_0)$, $e_k \in \mathrm{Ext}(B_{\mathcal{U}'})$ and $p \in \mathcal{N}'$.

Sketch of proof

1. Existence of a solution
   - The cost functional is coercive
   - Weak*-lower semicontinuity
   - The generalized Weierstrass theorem
2. Rewriting $\mathcal{V}$ as a constrained problem
   - Strict convexity of $\| \cdot - \boldsymbol{y}\|_2^2$
3. Removing N_0 constraints
   - Precise specification of $p \in \mathcal{N}'$
4. Reformulating the problem over $\mathcal{U}'$
5. Form of the extreme points
   - The general representer theorem over $\mathcal{U}'$

Part II: Supervised Learning with Sparsity Prior

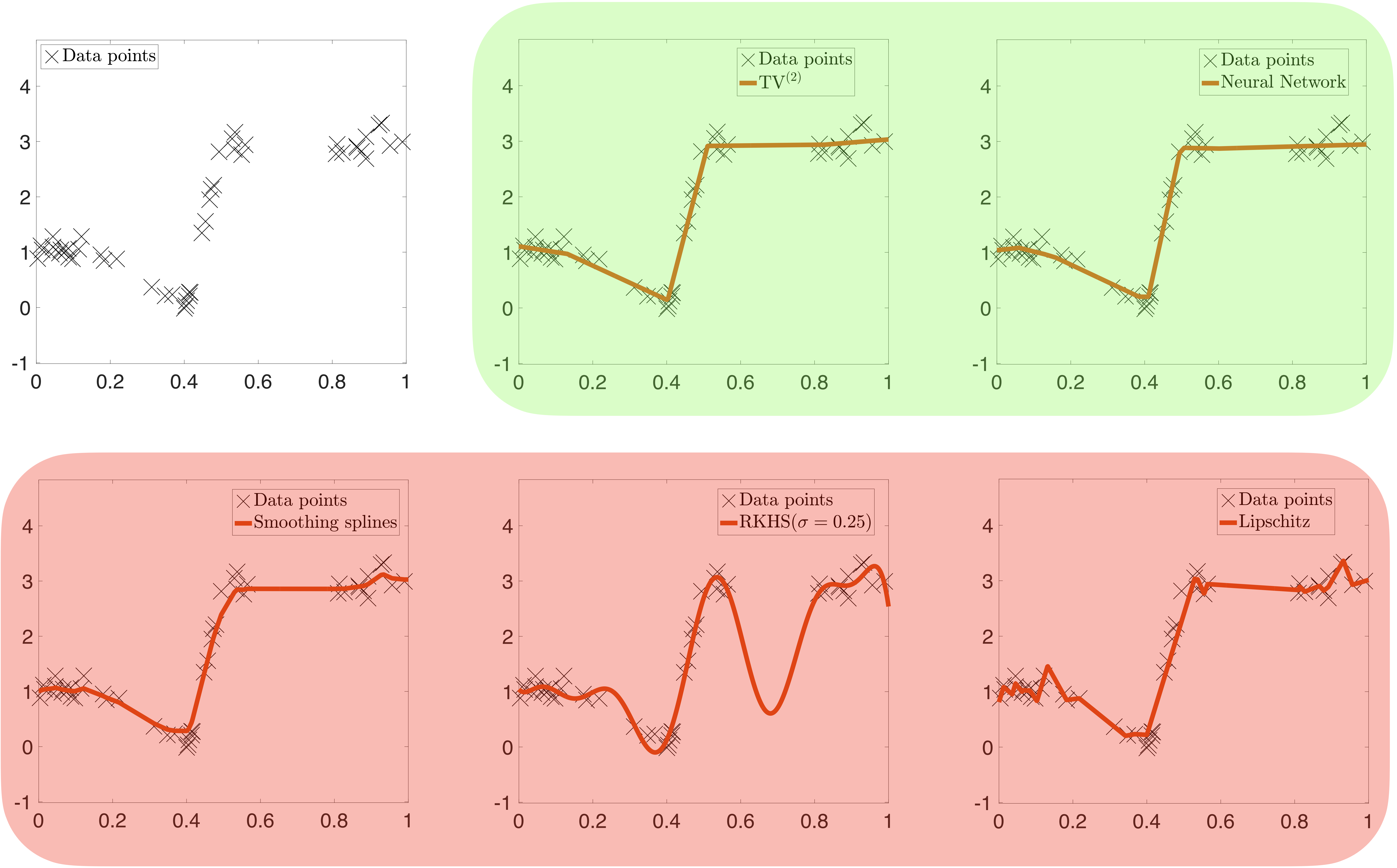
- Deriving regression schemes in the nonparametric setting

  1. Multi-kernel regression with sparse and adaptive kernels
  2. Learning univariate functions under joint sparsity and Lipschitz constraints
  3. Learning free-form activation functions of deep neural networks
  4. Learning multivariate continuous and piecewise linear functions

- Relevant publications

  - **S. Aziznejad**, M. Unser, "Multikernel regression with sparsity constraint," *SIAM Journal on Mathematics of Data Science, 2021*.
  - **S. Aziznejad**, T. Debarre, M. Unser, "Sparsest univariate learning models under Lipschitz constraint," *IEEE Open Journal of Signal Processing, 2022*.
  - **S. Aziznejad**, H. Gupta, J. Campos, M. Unser, "Deep neural networks with trainable activations and controlled Lipschitz constant," *IEEE Transactions on Signal Processing, 2020*.
  - 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*.
  - **S. Aziznejad**, M. Unser, "Duality mapping for Schatten matrix norms," *Numerical Functional Analysis and Optimization, 2021*.
  - **S. Aziznejad**, J. Campos, M. Unser, "Measuring complexity of learning schemes using Hessian-Schatten total variation," *ArXiv, 2021*.
  - J. Campos, **S. Aziznejad**, M. Unser, "Learning of continuous and piecewise-linear functions with Hessian total-variation regularization," *IEEE Open Journal of Signal Processing, 2022*.

Part II: Supervised Learning with Sparsity Prior



Part II: Supervised Learning with Sparsity Prior

- Deriving regression schemes in the nonparametric setting

  1. Multi-kernel regression with sparse and adaptive kernels

- Relevant publications

  - **S. Aziznejad**, M. Unser, "Multikernel regression with sparsity constraint," *SIAM Journal on Mathematics of Data Science*, *2021*.

Banach-Admissible Kernels

Johann Radon
(1887 – 1956)

- Recall: $\mathcal{M}(\mathbb{R}^d)$ is the space of finite Radon measures (Duval-Peyré '15) (Chizat-Bach '20)
  - $L_1(\mathbb{R}^d) \subseteq \mathcal{M}(\mathbb{R}^d)$ with $\|f\|_{L_1} = \|f\|_{\mathcal{M}}$ for any $f \in L_1(\mathbb{R}^d)$.
  - For any $\boldsymbol{a} = (a_n) \in \ell_1(\mathbb{Z})$:

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

- L: Linear shift-invariant (LSI) isomorphisms onto $\mathcal{M}(\mathbb{R}^d)$ (Unser et al. '17)

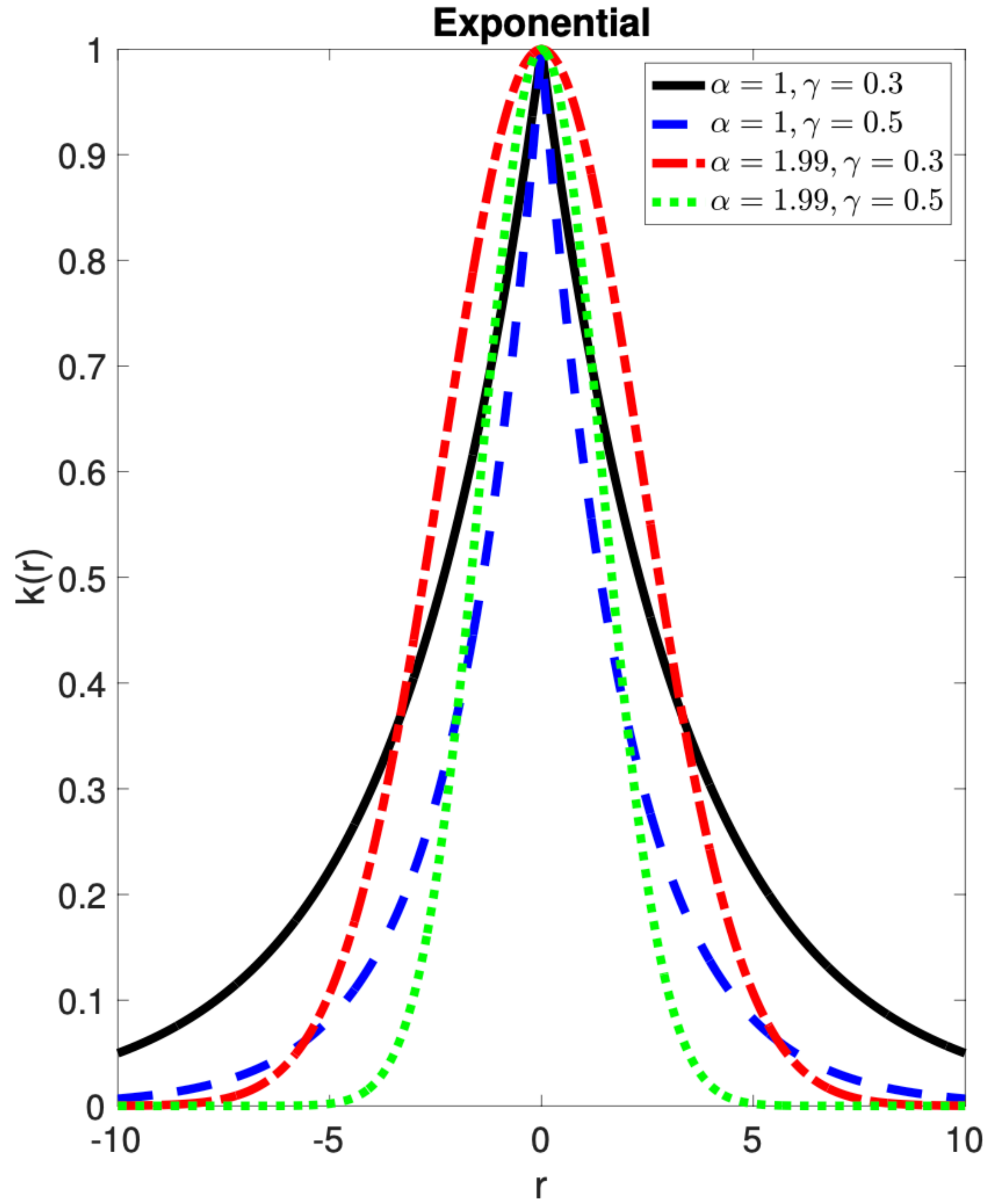
- Search space $\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) = \mathrm{L}^{-1}\left(\mathcal{M}(\mathbb{R}^d)\right)$
  - Banach structure: $\|f\|_{\mathcal{M}_{\mathrm{L}}} = \|\mathrm{L}\{f\}\|_{\mathcal{M}}$
  - Banach kernel: $\mathrm{k} = \mathrm{L}^{-1}\{\delta\} \in \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$
  - Extreme points of $B_{\mathcal{M}_{\mathrm{L}}}$: $\pm\mathrm{k}(\cdot - \boldsymbol{z}_0)$ for all $\boldsymbol{z}_0 \in \mathbb{R}^d$

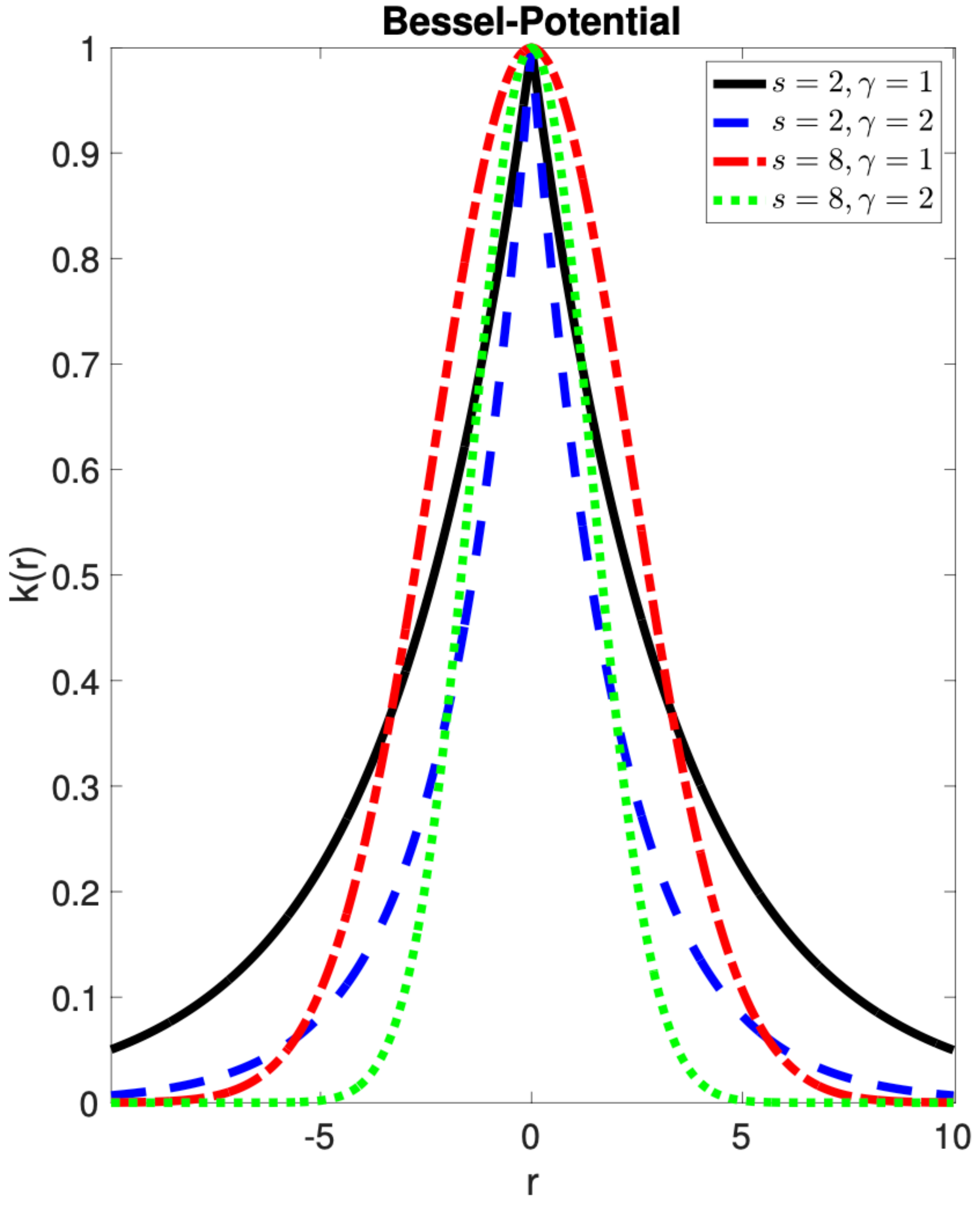
Banach-Admissible Kernels

Theorem [A.-Unser '21]

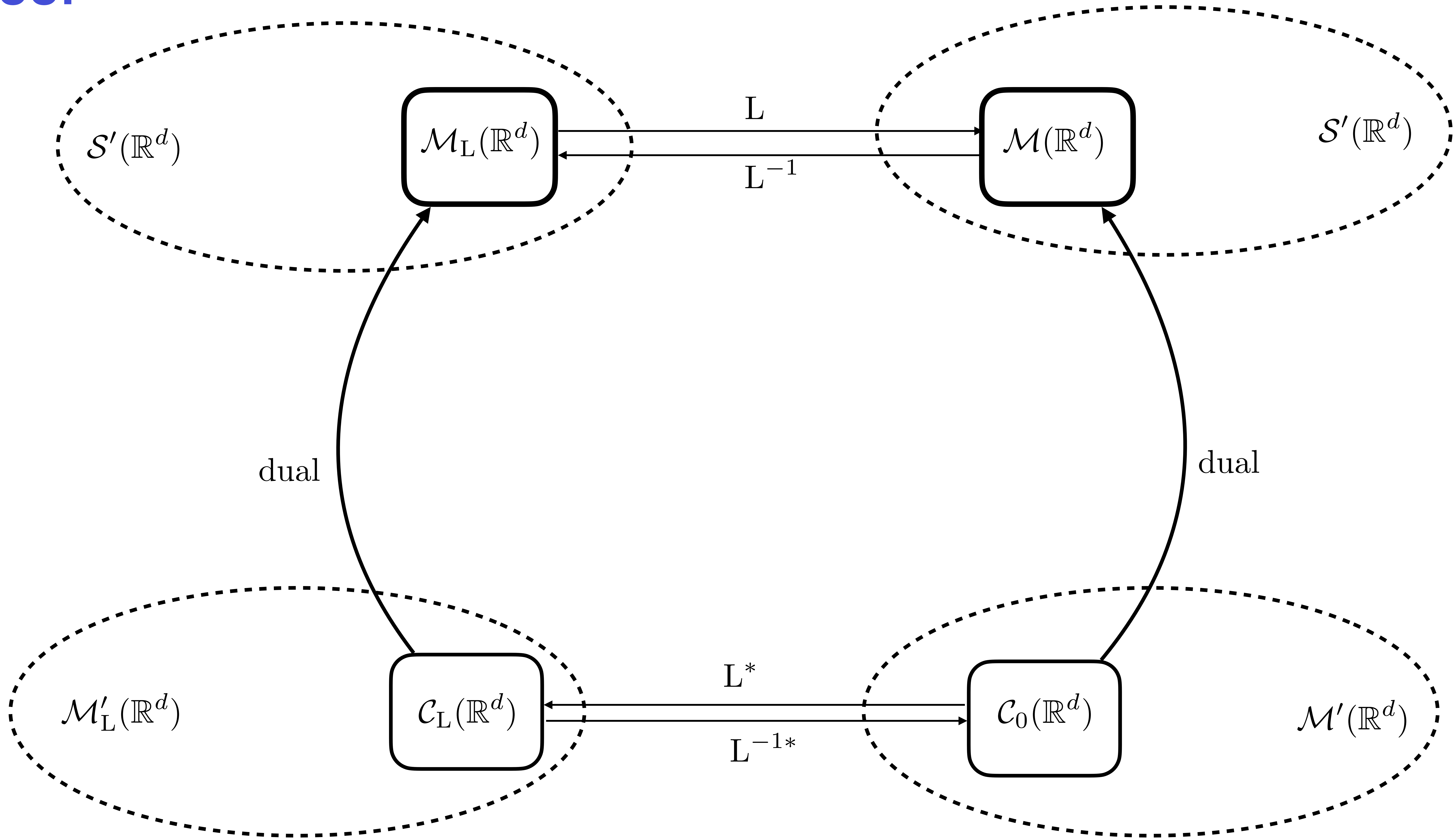
1. The LSI operator L is an isomorphism over $\mathcal{S}'(\mathbb{R}^d)$ if and only if the Fourier transform of its Banach kernel $\widehat{k}(\boldsymbol{\omega})$ is a smooth, nonvanishing, slowly growing, and heavy-tailed function of $\boldsymbol{\omega}$.

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





Proof


$\mathcal{S}'(\mathbb{R}^d)$
$\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$
L
L^{-1}
$\mathcal{M}(\mathbb{R}^d)$
$\mathcal{S}'(\mathbb{R}^d)$
dual
dual
$\mathcal{M}'_{\mathrm{L}}(\mathbb{R}^d)$
$\mathcal{C}_{\mathrm{L}}(\mathbb{R}^d)$
L^*
L^{-1*}
$\mathcal{C}_0(\mathbb{R}^d)$
$\mathcal{M}'(\mathbb{R}^d)$

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 $\mathrm{k}_{\boldsymbol{\mu}} = \sum_{n=1}^N \mu_n \mathrm{k}_n$

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

$$\min_{\substack{f_n \in \mathcal{M}_{\mathrm{L}_n}(\mathbb{R}^d), \\ f = \sum_{n=1}^N f_n}} \sum_{m=1}^M |f(\boldsymbol{x}_m) - y_m|^2 + \lambda \sum_{n=1}^N \|\mathrm{L}_n\{f_n\}\|_{\mathcal{M}},$$

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,

$$\boldsymbol{a}^* = (a_{n,l}^*) \in \arg\min_{\boldsymbol{a} \in \mathbb{R}^K} \sum_{m=1}^M \|\mathbf{G}\boldsymbol{a} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{a}\|_{\ell_1}$$

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

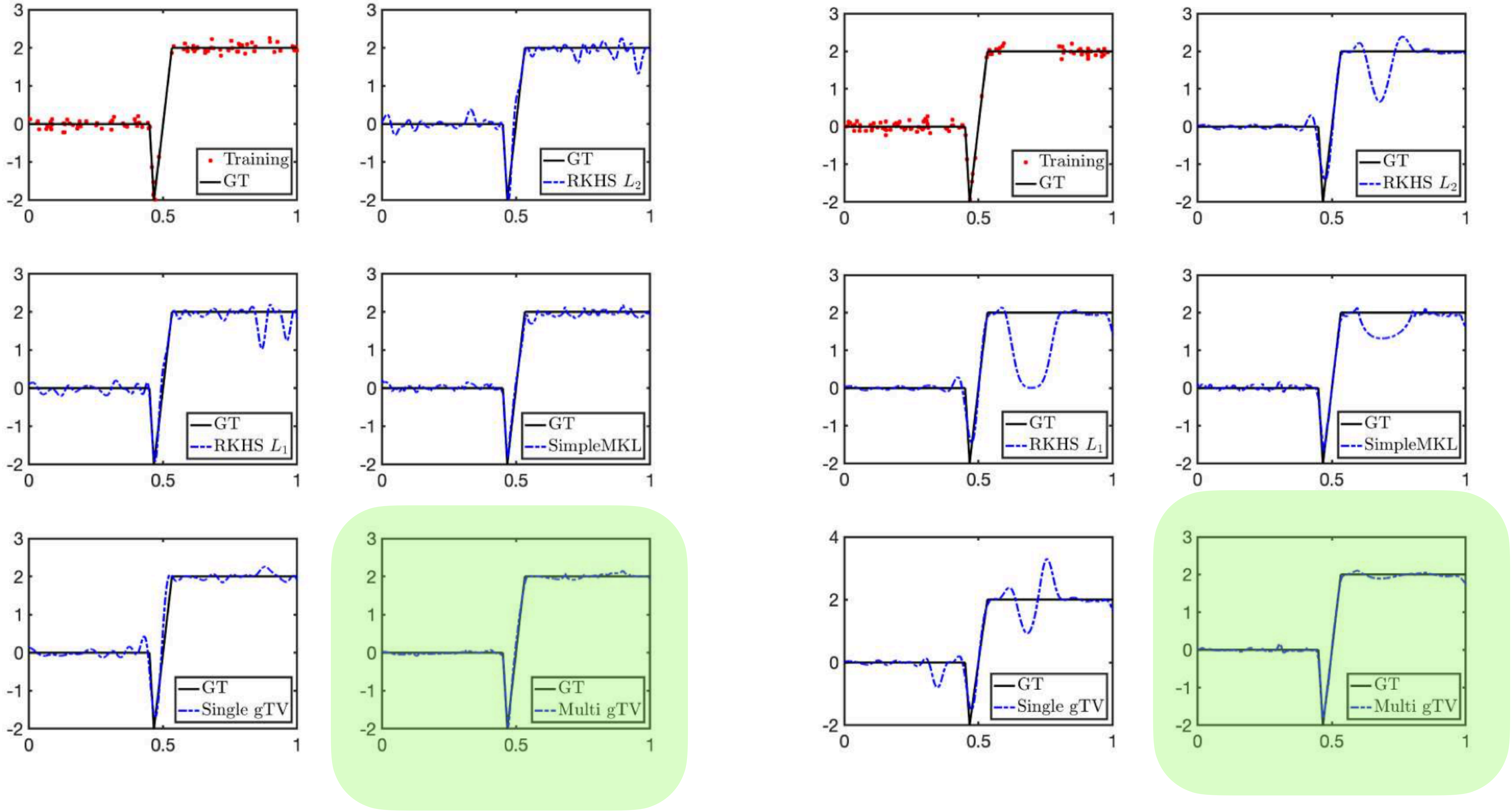
Sketch of proof

1. Search space: $\mathcal{X}' = \prod_{n=1}^N \mathcal{M}_{\mathrm{L}_n}(\mathbb{R}^d)$
2. Measurements: $\nu_m(f_1, \dots, f_N) = \sum_{n=1}^N f_n(\boldsymbol{x}_m)$
3. The representer theorem for $\mathcal{X}'$

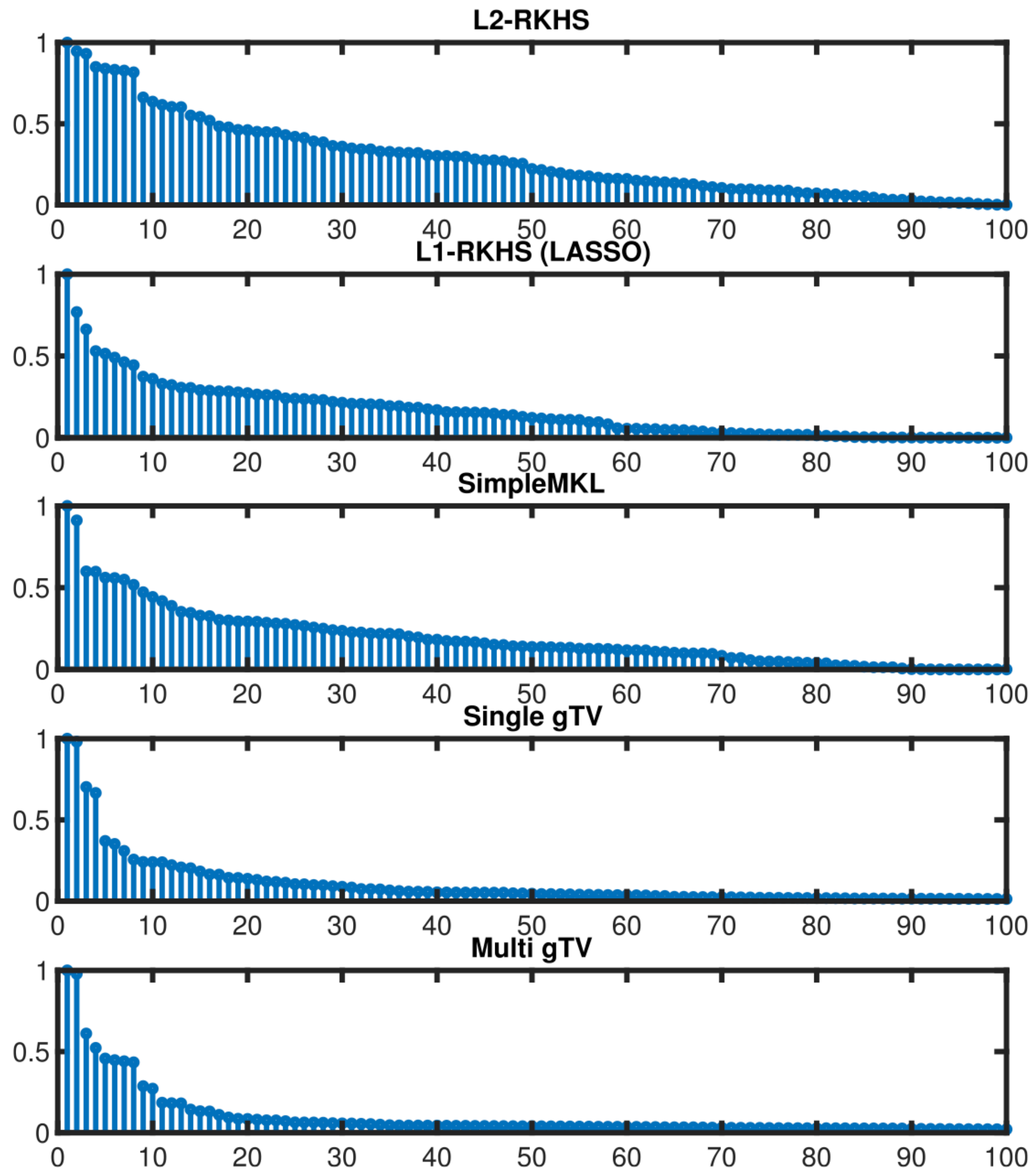
Practical outcomes

1. $K \leq M$: The upper-bound is independent of N
2. Adaptive expansion: both in shapes and locations
3. Sparse expansion: ℓ_1 penalty on kernel coefficients
4. In low dimensions: Grid-based methods + FISTA

Numerical Examples



(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**

Part II: Supervised Learning with Sparsity Prior

- Deriving regression schemes in the nonparametric setting

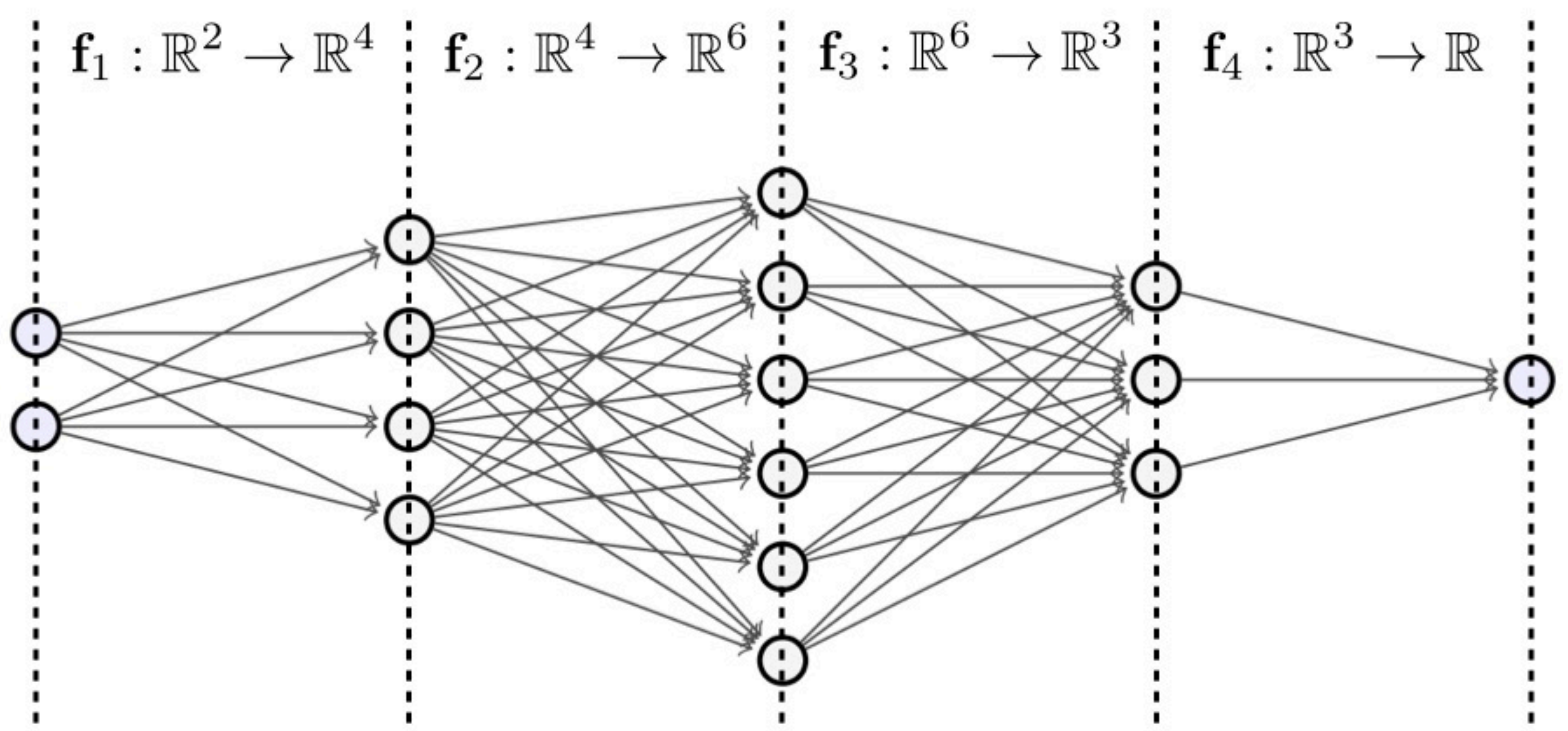
2. Learning univariate functions under joint sparsity and Lipschitz constraints

3. Learning free-form activation functions of deep neural networks

- Relevant publications

- **S. Aziznejad**, T. Debarre, M. Unser, "Sparsest univariate learning models under Lipschitz constraint," *IEEE Open Journal of Signal Processing, 2022*.
- **S. Aziznejad**, H. Gupta, J. Campos, M. Unser, "Deep neural networks with trainable activations and controlled Lipschitz constant," *IEEE Transactions on Signal Processing, 2020*.
- 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*.

Feed-Forward Deep Neural Networks



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

- Composition of “simple” vector-valued mappings

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

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

  - 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 $\boldsymbol{\sigma}_l : \mathbb{R}^{N_l} \to \mathbb{R}^{N_l} \quad (x_1, \ldots, x_{N_l}) \mapsto (\sigma_{1,l}(x_1), \sigma_{2,l}(x_2), \ldots, \sigma_{N_l,l}(x_{N_l}))$
  - Alternative representation $\mathbf{f}_l = \boldsymbol{\sigma}_l \circ \mathbf{W}_l$

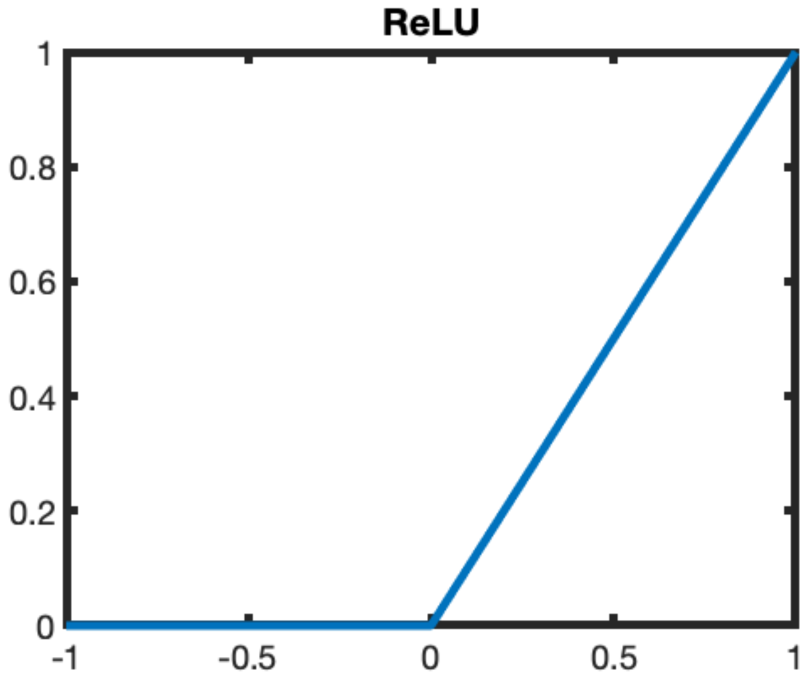
- Fixed-shape nonlinearities $\sigma_{n,l}(x) = \sigma(x - b_{n,l})$

Activation Functions

- Fixed activation functions: ReLU, LReLU

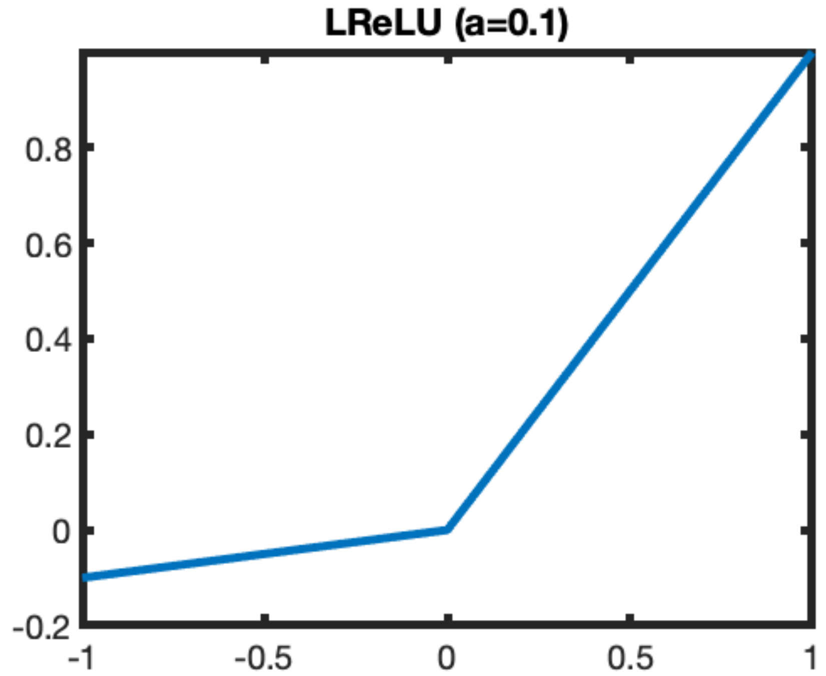
$$\mathrm{ReLU}(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

(Glorot *et al.* '11)



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

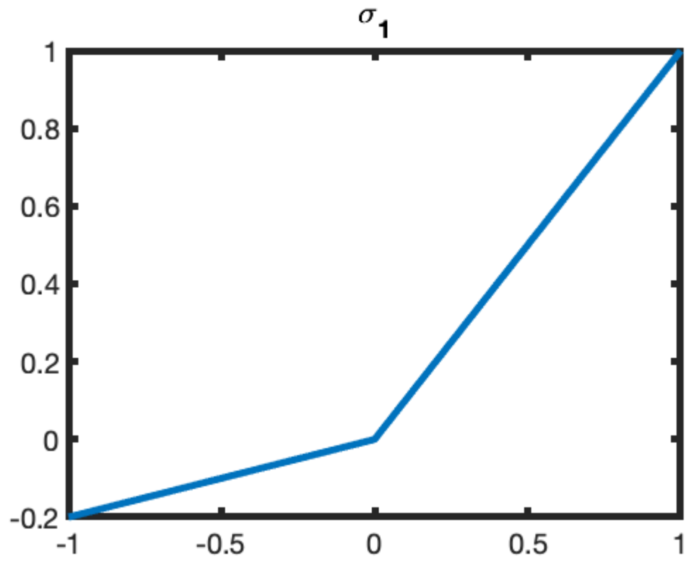
(Maas *et al.* '13)

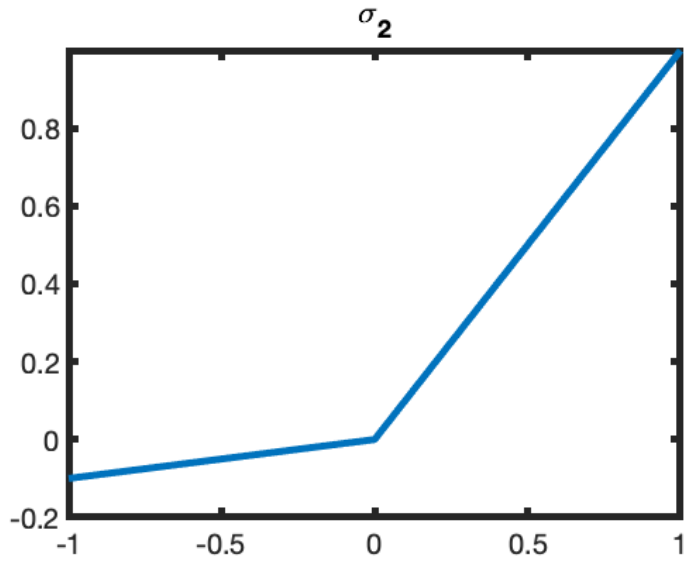


- Parametric activation functions

PReLU: Learn the negative slope

(He *et al.* '15)

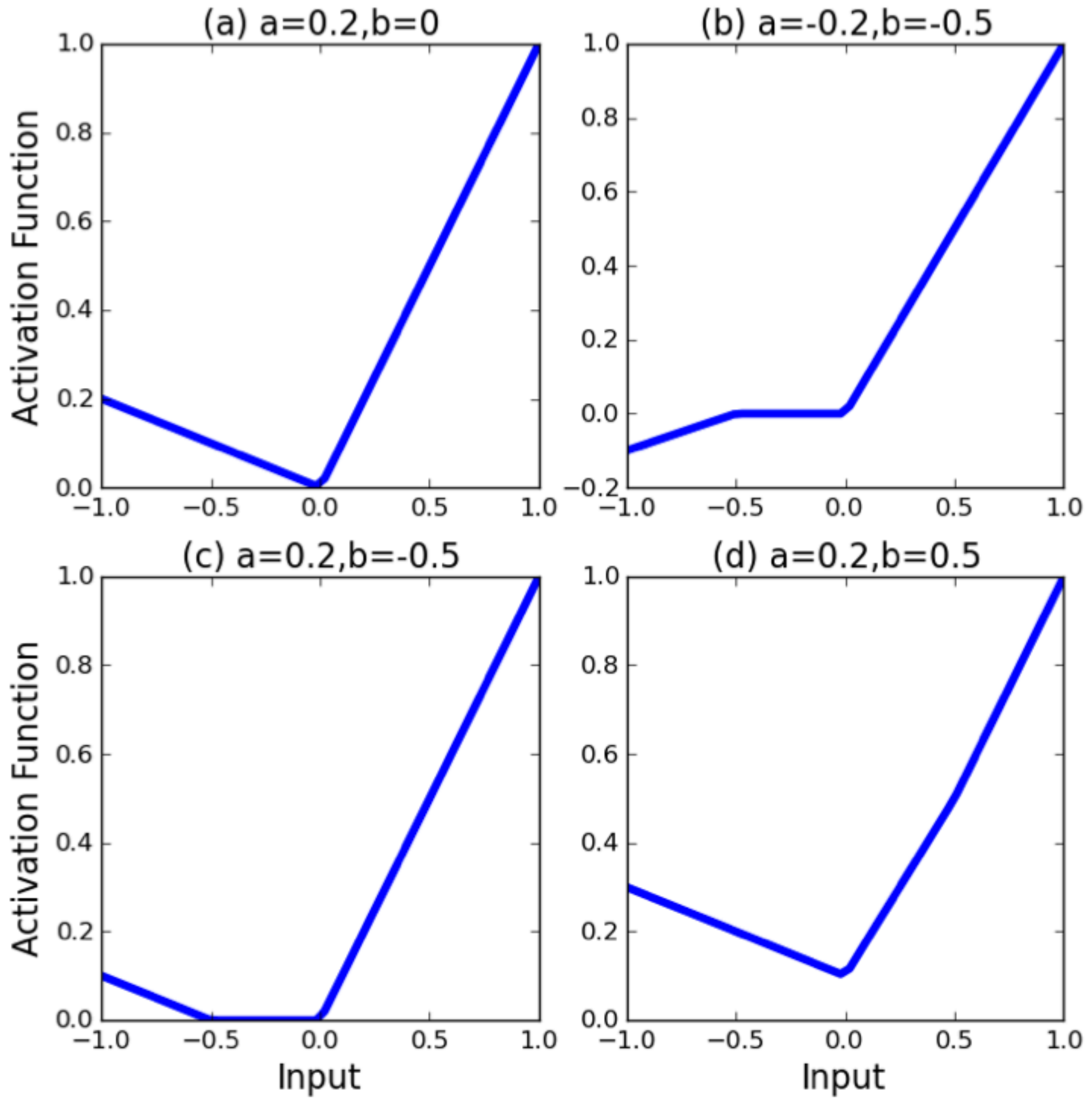




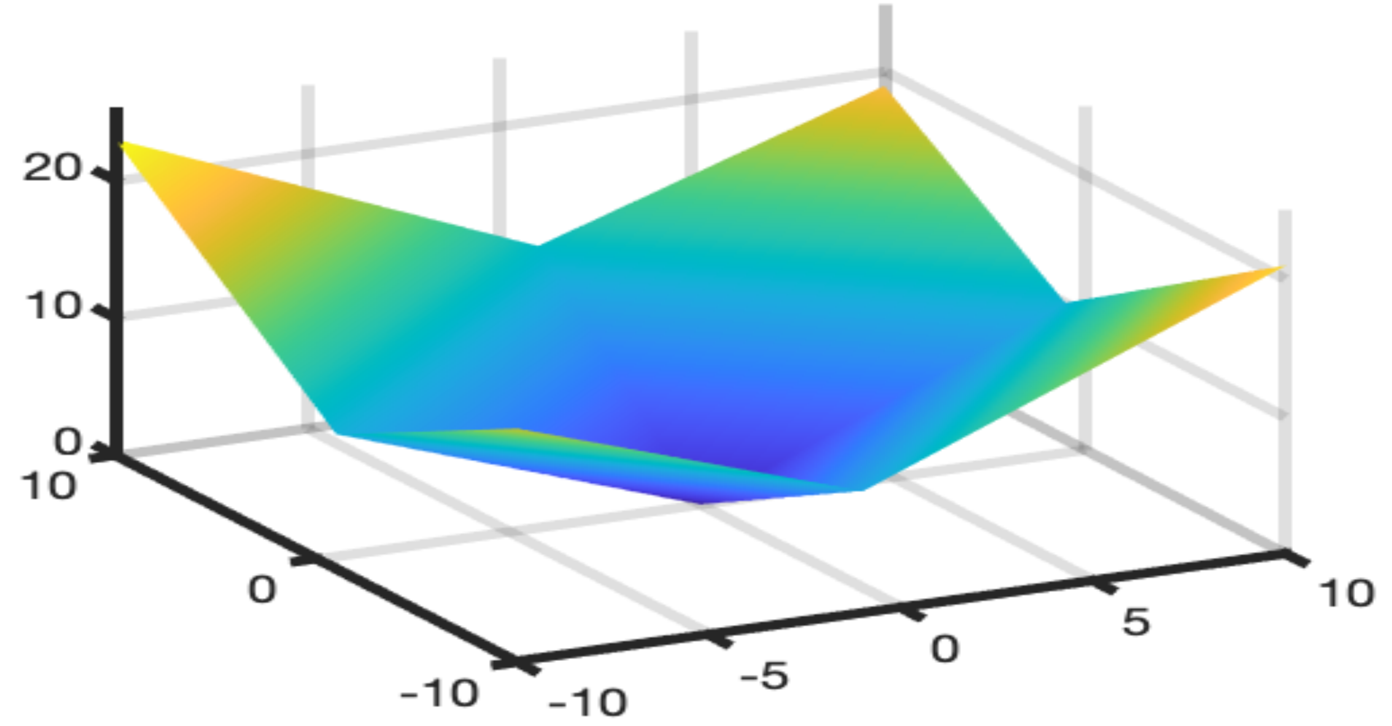
Adaptive Piecewise Linear

(Agostinelli *et al.* '15)

- Linear spline
- ℓ_2 regularization
- < 10 knots

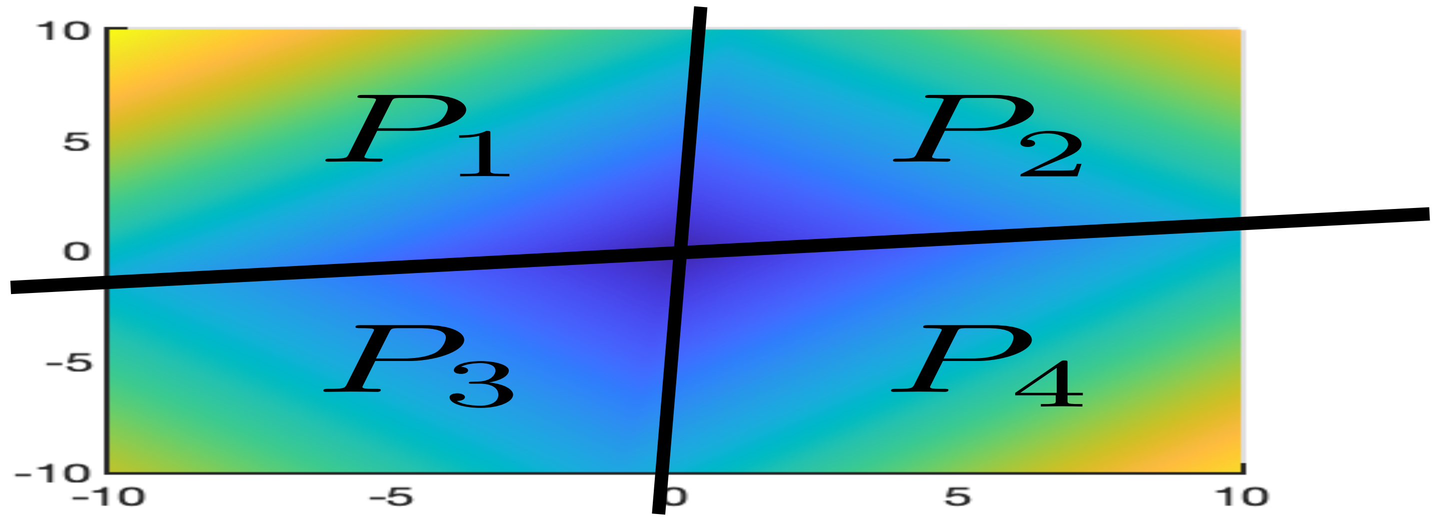


CPWL Structure of ReLU Neural Networks



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

- Continuous and Piecewise-Linear (CPWL) Functions
  - $f \in \mathcal{C}(\mathbb{R}^d)$
  - $\exists (P_n)_{n=1}^N : \mathbb{R}^d = P_1 \sqcup \cdots \sqcup P_N$ and $f|_{P_n}$ is affine for $n = 1, \ldots, N$.

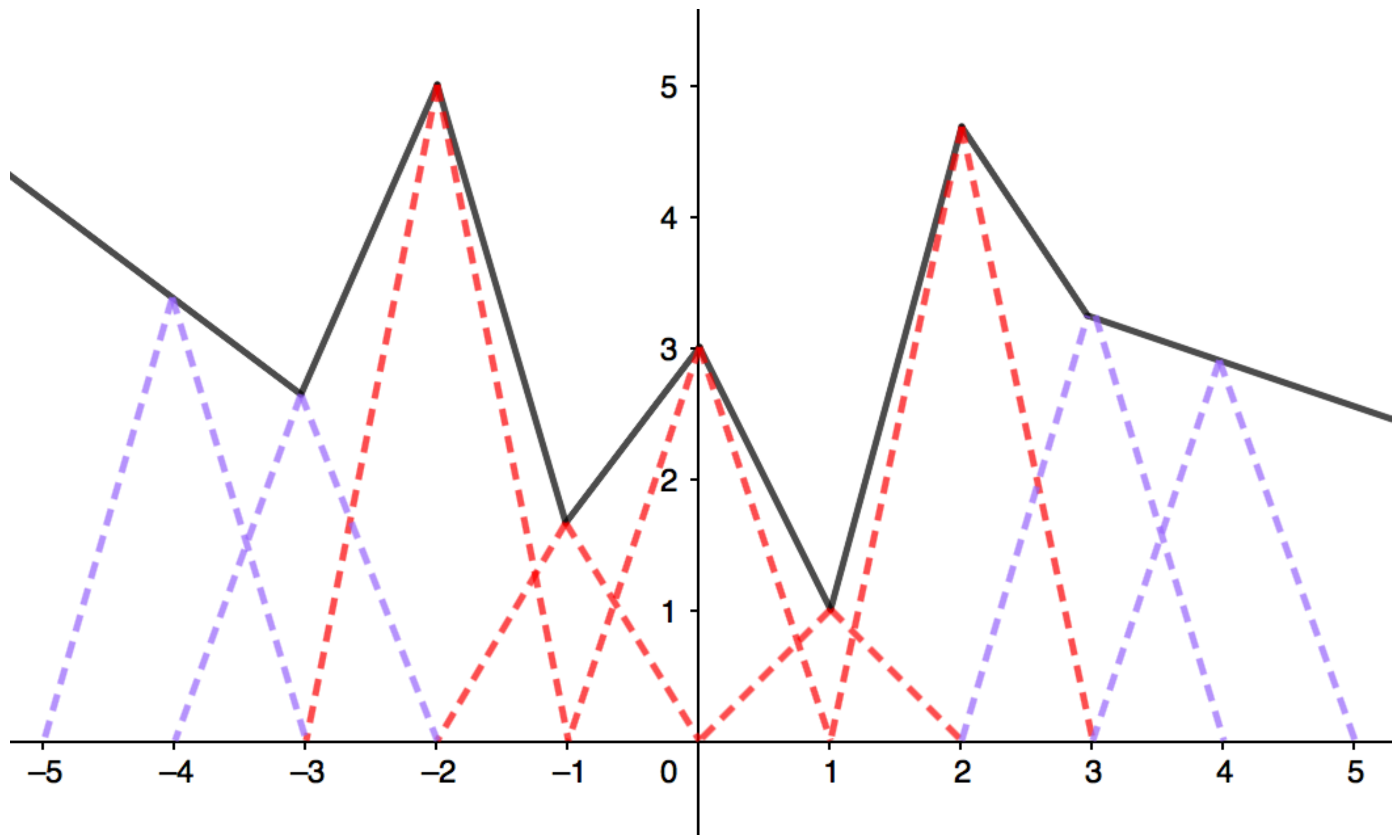


- CPWL structure of ReLU DNNs
  - In 1D: CPWL $\iff$ Linear spline
  - Linear combination of CPWL functions $\Rightarrow$ CPWL
  - Composition of two CPWL $\Rightarrow$ CPWL

  $\Rightarrow$ linear spline DNNs are CPWL.

- Converse: CPWL functions can be represented by ReLU DNNs. (Arora *et al.* '18)

Free-Form Activation Functions



Deep Splines!

- Principled design:
  - Preserves CPWL structure of DNNs
  - Promotes sparse activation functions
  - Controls the global Lipschitz regularity of the network (Antun *et al.* '20)
  - Efficient implementation that makes it scalable in time and memory

- Deep splines: a functional framework for learning activation functions

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

Part II: Supervised Learning with Sparsity Prior

- Deriving regression schemes in the nonparametric setting

  2. Learning univariate functions under joint sparsity and Lipschitz constraints

- Relevant publications

  - **S. Aziznejad**, T. Debarre, M. Unser, "Sparsest univariate learning models under Lipschitz constraint," *IEEE Open Journal of Signal Processing*, *2022*.

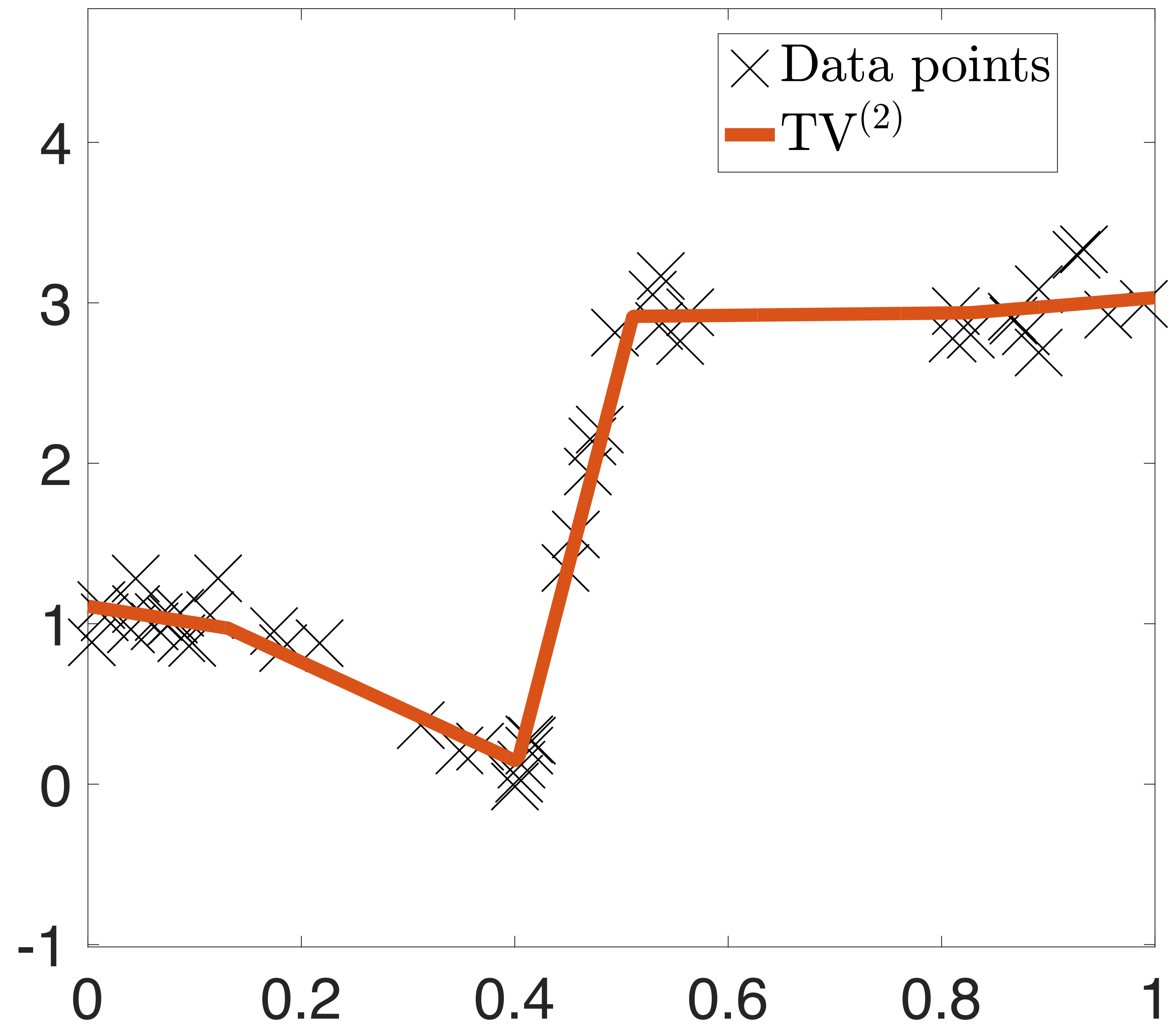
1D Regression with Sparsity

- Simple observation:

Sparsity promoting!

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

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



- $\mathcal{V}_{\mathrm{TV}^{(2)}}$ contains linear spline solutions with at most $(M-2)$ knots.

(Gupta *et al.* '18) (Unser *et al.* '17)

- Efficient method for finding the sparsest linear spline solution

(Debarre *et al.* '22)

1D Regression: 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. '22, simplified]
The solution set

$$\mathcal{V}_{\mathrm{Lip}} = \underset{f \in \mathrm{Lip}(\mathbb{R})}{\arg\min} \sum_{m=1}^{M} |f(x_m) - y_m|^2 + \lambda L(f)$$

is nonempty, convex, and weak*-compact. Moreover, there exists a unique vector $\boldsymbol{z} = (z_m) \in \mathbb{R}^M$ such that

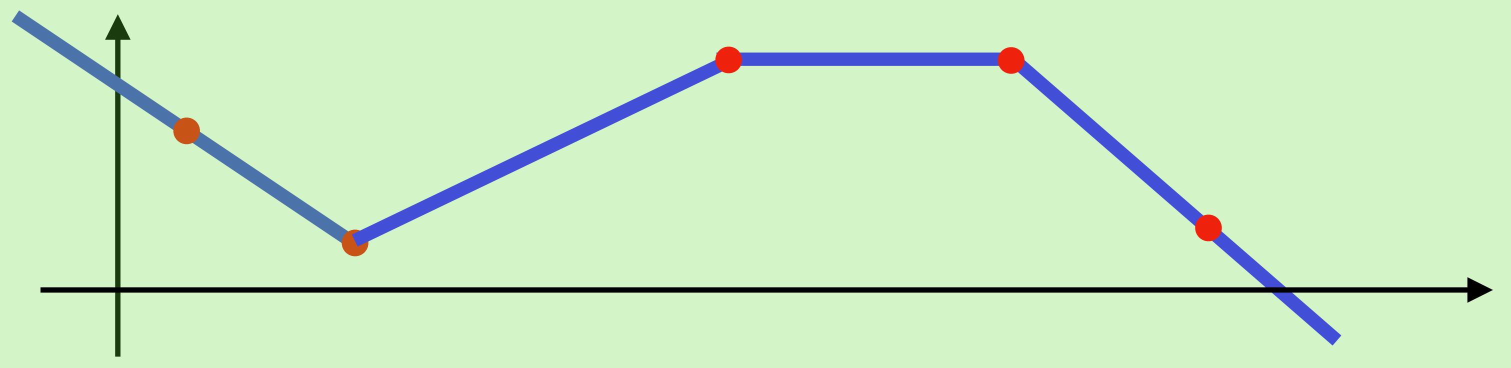
$$\mathcal{V}_{\mathrm{Lip}} = \left\{ f \in \mathrm{Lip}(\mathbb{R}) : L(f) = \max_{m \neq n} \left| \frac{z_m - z_n}{x_m - x_n} \right|, \forall m : f(x_m) = z_m \right\}$$

Corollary: The solution set $\mathcal{V}_{\mathrm{Lip}}$ contains linear splines.

Proof. Take the canonical linear spline interpolator of $\{(x_m, z_m)\}_{m=1}^{M}$.

Sketch of proof

1. Topological structure of $\mathcal{V}_{\mathrm{Lip}}$
   - Finding the predual of $\mathrm{Lip}(\mathbb{R})$
   - Weak*-continuity of sampling
   - Representer theorem for seminorms
2. Existence of $\boldsymbol{z}$
   - Strict convexity of $\|\cdot - \boldsymbol{y}\|_2^2$
3. $f_{\mathrm{cano}} \in \mathcal{V}_{\mathrm{Lip}}$



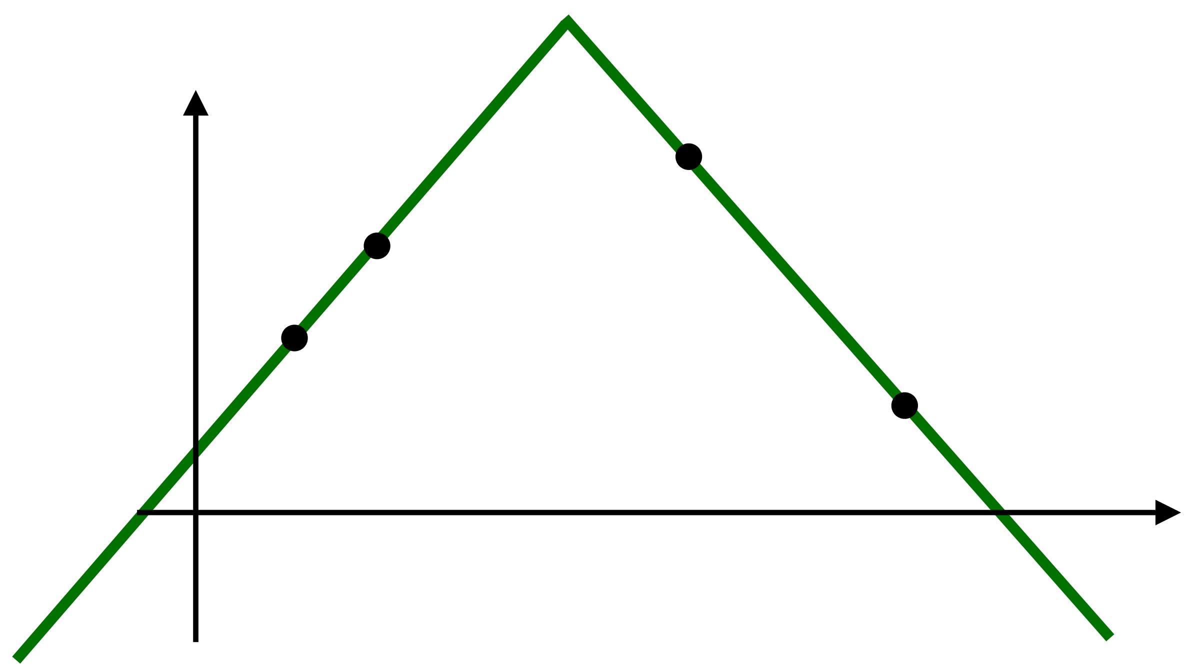
How to find the sparsest solution?

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

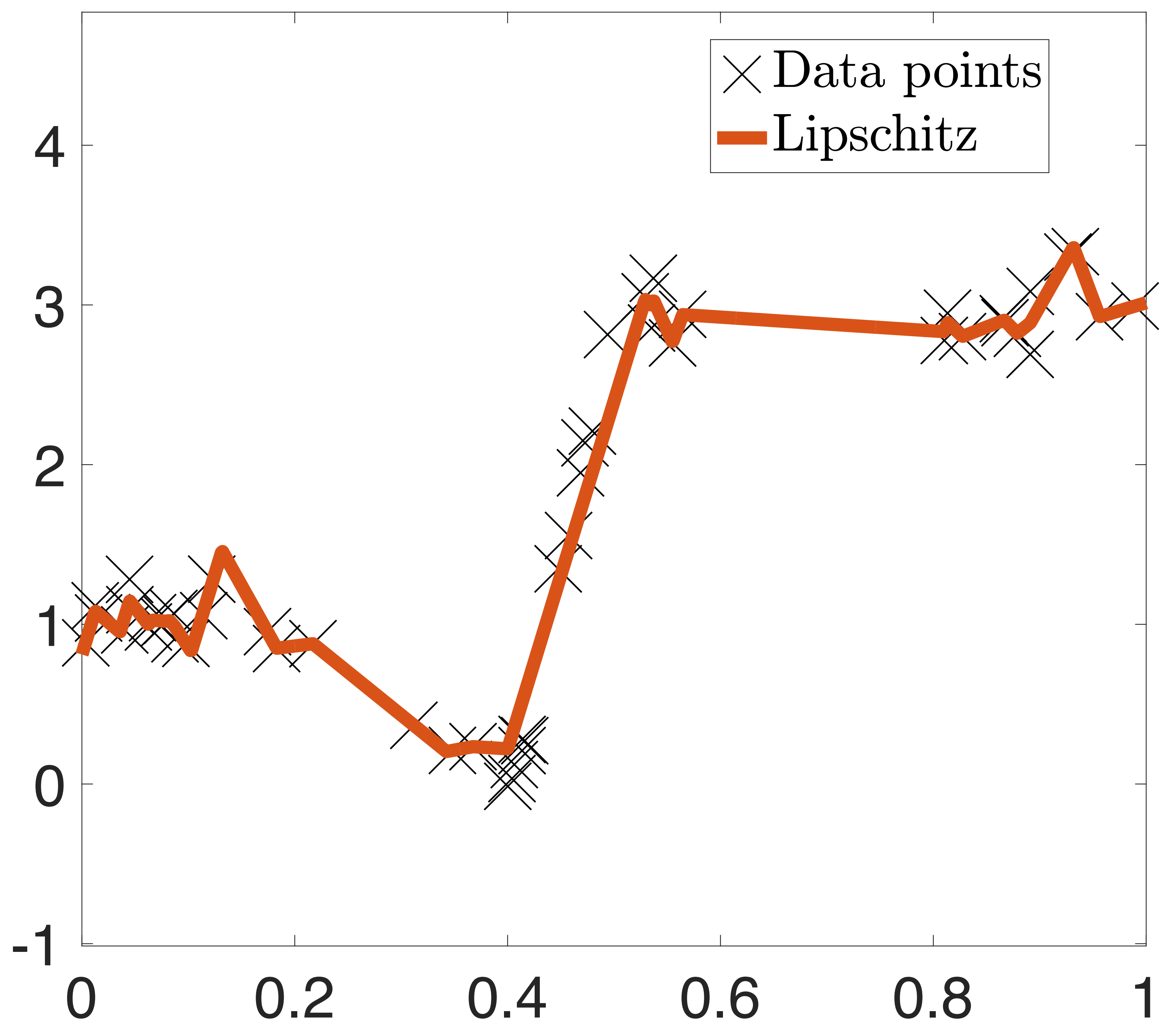
  - Using proximal methods (*e.g. ADMM*), solve the minimization

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

  - 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: Sparse + Lipschitz

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

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

Theorem [A. et al. '21]

- $\mathcal{V}_{\mathrm{hyb}}$: nonempty, convex and weak*-compact subset of $\mathrm{BV}^{(2)}(\mathbb{R})$
- Extreme points of $\mathcal{V}_{\mathrm{hyb}}$: linear splines with $K \leq M$ knots.
- Let us denote by $\boldsymbol{\theta}$, the parameter vector of the shallow ReLU network $f_{\boldsymbol{\theta}} : \mathbb{R} \to \mathbb{R}$ with two layers and skip connections. Consider the minimization problem

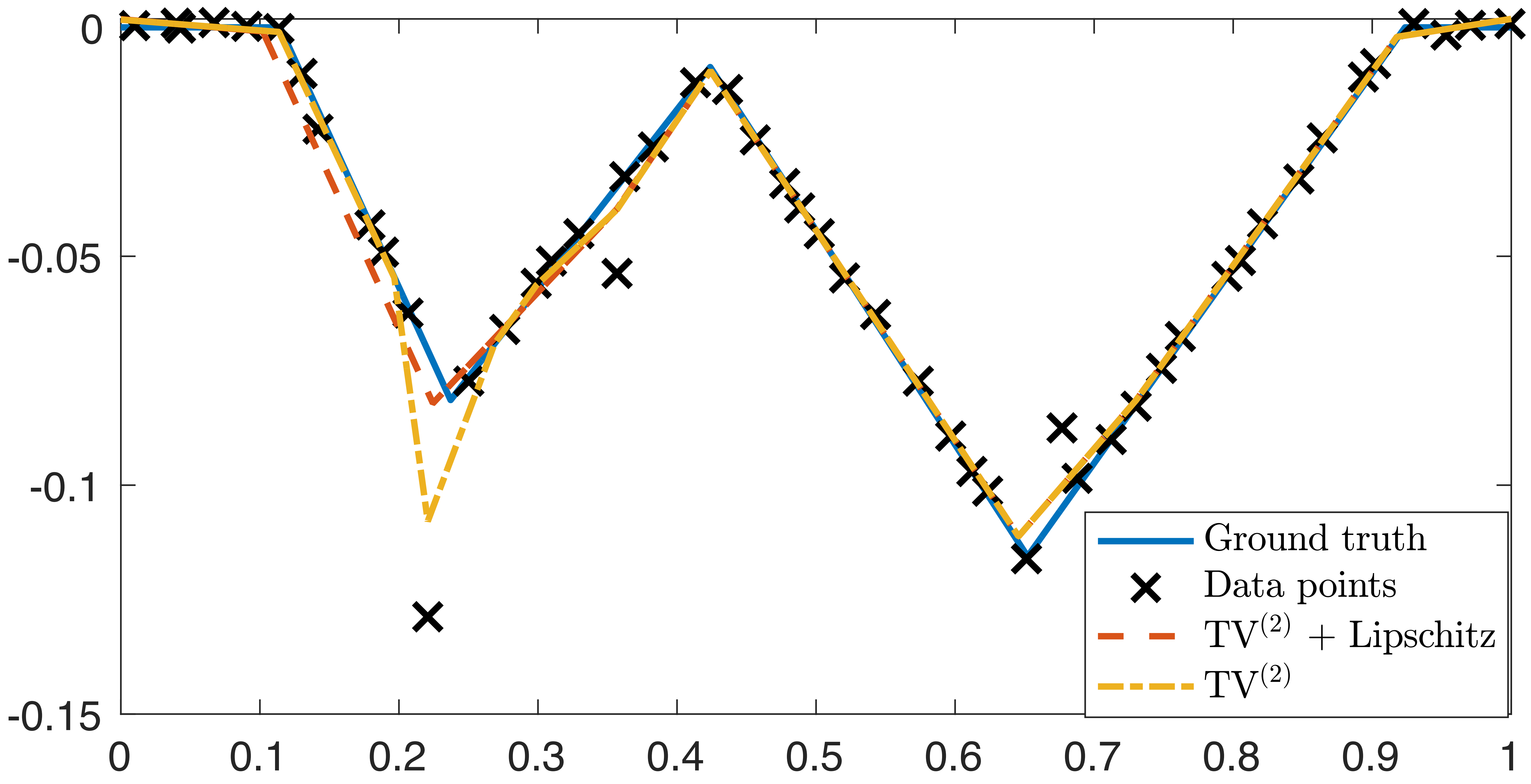
$$\mathcal{V}_{NN} = \arg\min_{\boldsymbol{\theta}} \sum_{m=1}^{M} |f_{\boldsymbol{\theta}}(x_m) - y_m|^2 + \lambda R(\boldsymbol{\theta}), \quad \text{s.t.} \quad L(f_{\boldsymbol{\theta}}) \leq \bar{L},$$

where $R(\boldsymbol{\theta})$ denotes weight decay regularization. Then the mapping $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}} : \mathcal{V}_{\mathrm{NN}} \to \mathcal{V}_{\mathrm{hyb}} \cap \mathrm{CPWL}$ is a bijection. (Parhi-Nowak *'21*)(Savarese et al. '19)

Sketch of proof

1. Topological structure of $\mathcal{V}_{\mathrm{hyb}}$
   - Weak*-closedness of the Lipschitz ball
   - Representer theorem for seminorms
2. Extreme points of $\mathcal{V}_{\mathrm{hyb}}$
   - $\mathcal{V}_{\mathrm{hyb}} = \mathcal{V}_{\mathrm{TV}^{(2)}}$ (informal)
3. Bijection with $\mathcal{V}_{\mathrm{NN}}$
   - Homogeneity of ReLU: $(2x)_+ = 2(x)_+$
   - $\mathrm{R}(\boldsymbol{\theta}^*) = \mathrm{TV}^{(2)}(f_{\boldsymbol{\theta}^*})$

1D Regression: Sparse + Lipschitz



Removing outliers!

Part II: Supervised Learning with Sparsity Prior

- Deriving regression schemes in the nonparametric setting

3. Learning free-form activation functions of deep neural networks

- Relevant publications

- **S. Aziznejad**, H. Gupta, J. Campos, M. Unser, "Deep neural networks with trainable activations and controlled Lipschitz constant," *IEEE Transactions on Signal Processing, 2020*.
- 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*.

Deep Splines Representer Theorem

- $L(f) \leq \|f\|_{\mathrm{BV}^{(2)}} = \mathrm{TV}^{(2)}(f) + |f(0)| + |f(1)|$
- $\boldsymbol{\sigma} = (\sigma_n) \in \mathrm{BV}^{(2)}(\mathbb{R})^N \Rightarrow \|\boldsymbol{\sigma}\|_{\mathrm{BV}^{(2)}} = \sum_{n=1}^N \|\sigma_n\|_{\mathrm{BV}^{(2)}}$

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 $\mathbf{f}_{\mathrm{deep}} : (\mathbb{R}^{N_0}, \|\cdot\|_2) \to (\mathbb{R}^{N_L}, \|\cdot\|_2)$ is upper-bounded by

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

Sketch of proof

1. Lipschitz constant of an activation function < TV2
2. For a layer: Hölder's ineqaulity
3. For the network: Product bound

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.

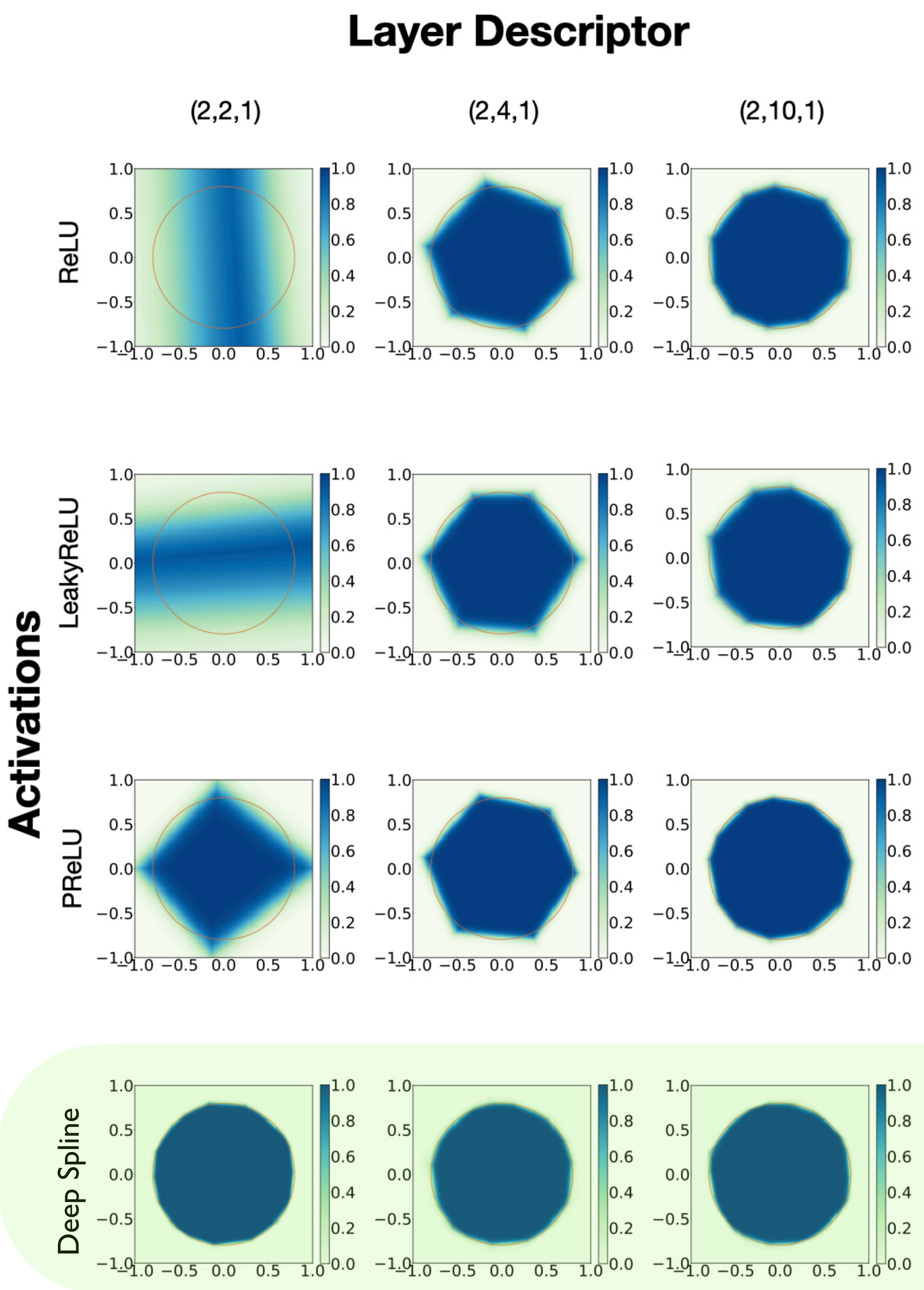
Moreover, any local minima of the above problem satisfies

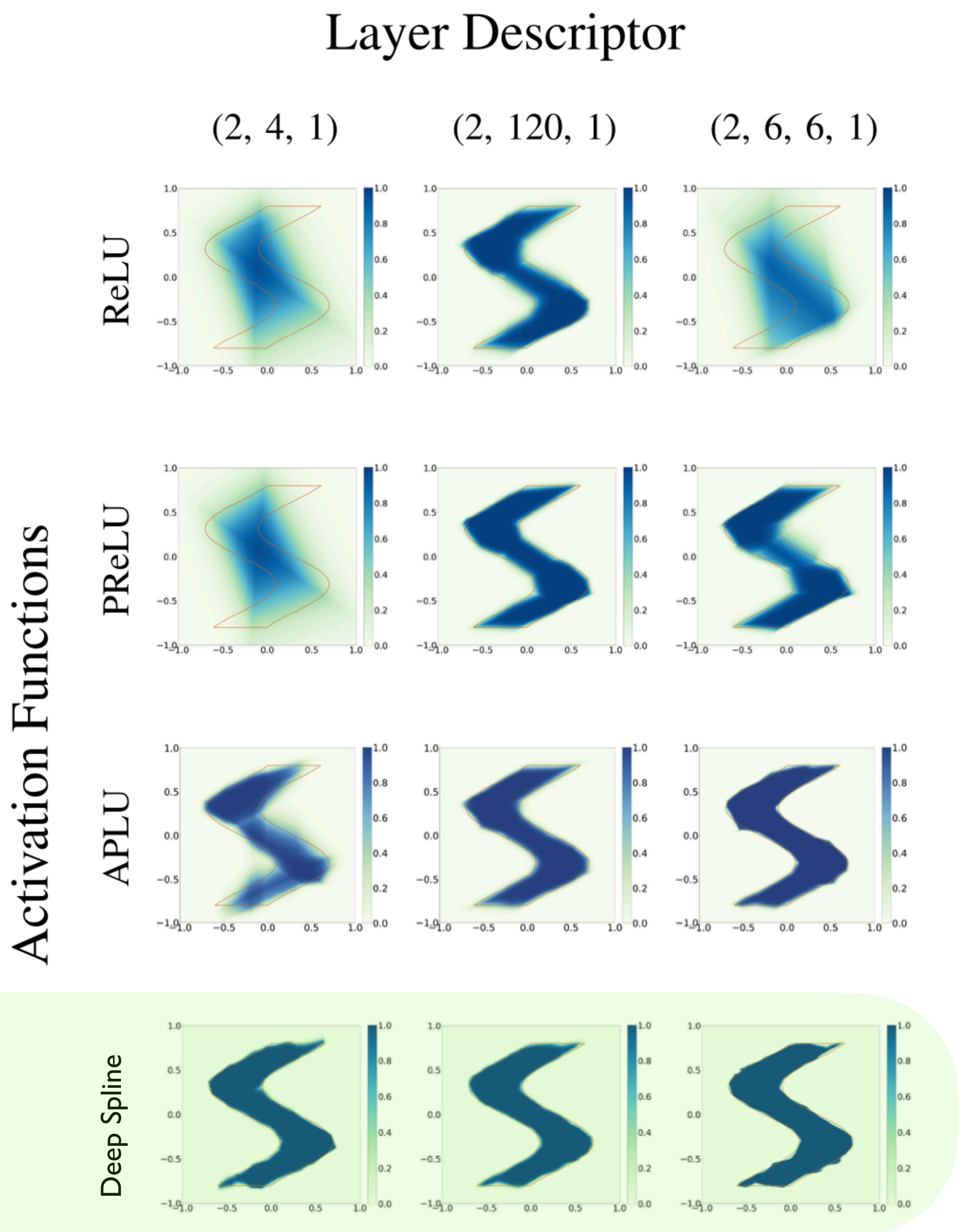
$$\lambda_l \|\boldsymbol{\sigma}_l\|_{\mathrm{BV}^{(2)}} = 2\mu_{l+1} \|\mathbf{W}_{l+1}\|_F^2, \quad l = 1, \dots, L-1.$$

Sketch of proof

1. Existence: Lipschitz-continuity of the activations
2. Form of the activation functions:
   - Fix an arbitrary solution
   - Define a 1D problem per activation function
   - Show the equivalence to the training of the neural network.
3. Optimality condition:
   - Homogeneity of TV2-regularization
   - AM-GM type inequality

Example


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 Spline

Part II: Supervised Learning with Sparsity Prior

- Deriving regression schemes in the nonparametric setting

    4. Learning multivariate continuous and piecewise linear functions

- Relevant publications

- **S. Aziznejad**, M. Unser, "Duality mapping for Schatten matrix norms," *Numerical Functional Analysis and Optimization*, *2021*.
- **S. Aziznejad**, J. Campos, M. Unser, "Measuring complexity of learning schemes using Hessian-Schatten total variation," *ArXiv*, *2021*.
- J. Campos, **S. Aziznejad**, M. Unser, "Learning of continuous and piecewise-linear functions with Hessian total-variation regularization," *IEEE Open Journal of Signal Processing*, *2022*.

CPWL Functions Revisited

- Recall: ReLU DNNs = Deep splines = CPWL family

- Goal: Learning CPWL mappings directly from the data

$$\min_{f \in \mathcal{F}(\mathbb{R}^d)} \sum_{m=1}^{M} |f(\boldsymbol{x}_m) - y_m|^2 + \lambda \mathcal{R}(f)$$

- Search space: $f \in \mathcal{F}(\mathbb{R}^d) \Leftrightarrow \mathcal{R}(f) < +\infty$
- Regularization: Sparsity-promoting, CPWL-promoting

In d=1: TV-2!

- Hessian of CPWL functions has Hausdorff dimension $= (d-1)$
- Schatten norms promote low-rank matrices
- Total-variation promotes sparsity in the space of measures

Hessian-Schatten Total Variation (HTV)

- Informal definition $\mathrm{HTV}_p(f) = \int_{\mathbb{R}^d} \|\mathrm{H}\{f\}(\boldsymbol{x})\|_{S_p} \mathrm{d}\boldsymbol{x}$ Not suitable for CPWL functions!

Hessian-Schatten Total Variation

Definition [A. et al. '21]

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

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

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

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

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

Sketch of proof

Item 1:

(I) LHS $\leq$ RHS

- Hölder's inequality

(II) $\forall \epsilon > 0$: LHS $\geq$ RHS $- \epsilon$

- Lusin's theorem
- Duality mapping of Schatten norms [A.-Unser'21]

Item 2:

(I) Assuming general conditions

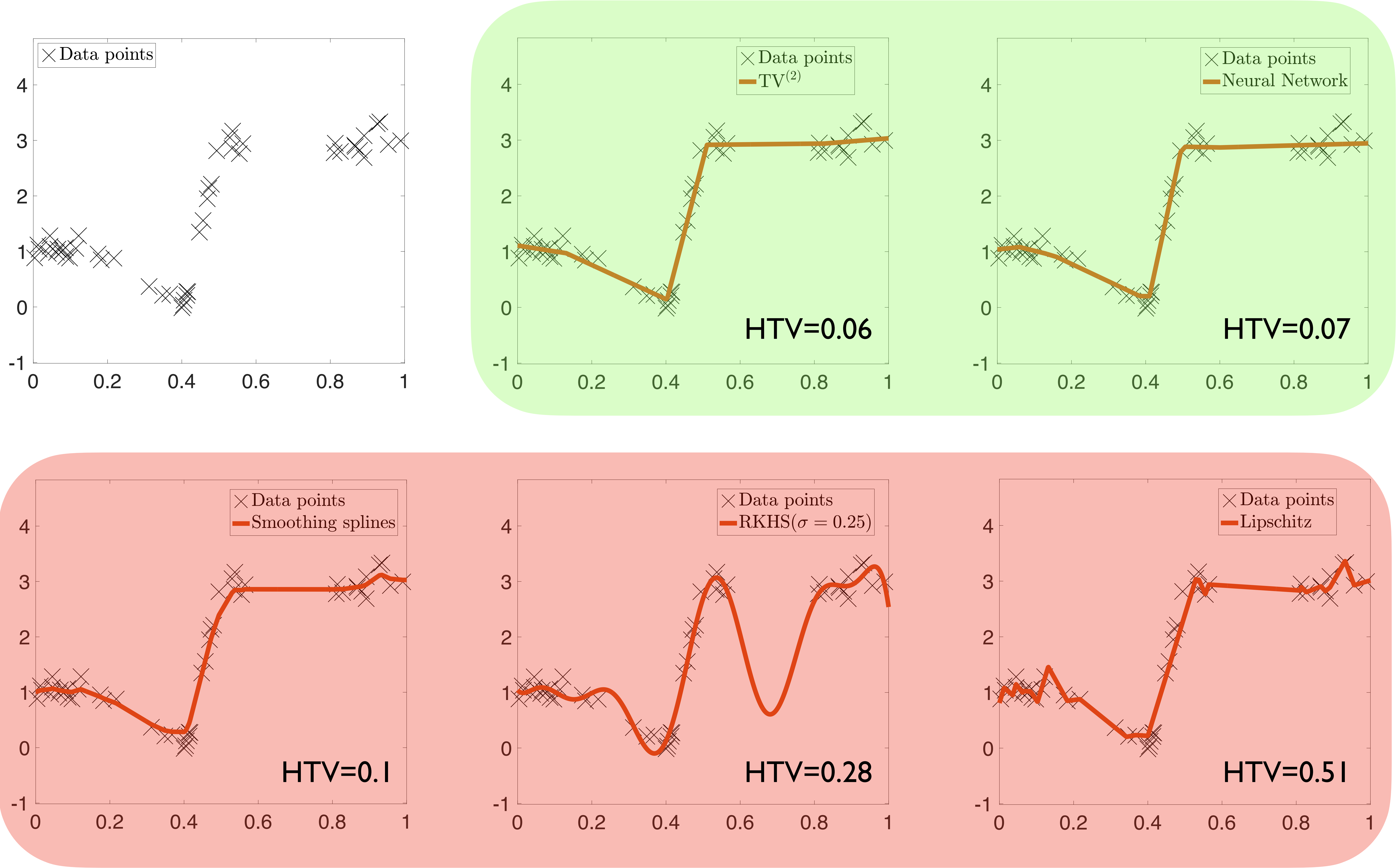
- invariance properties of the HTV

(II) Explicit computation of the Hessian measure

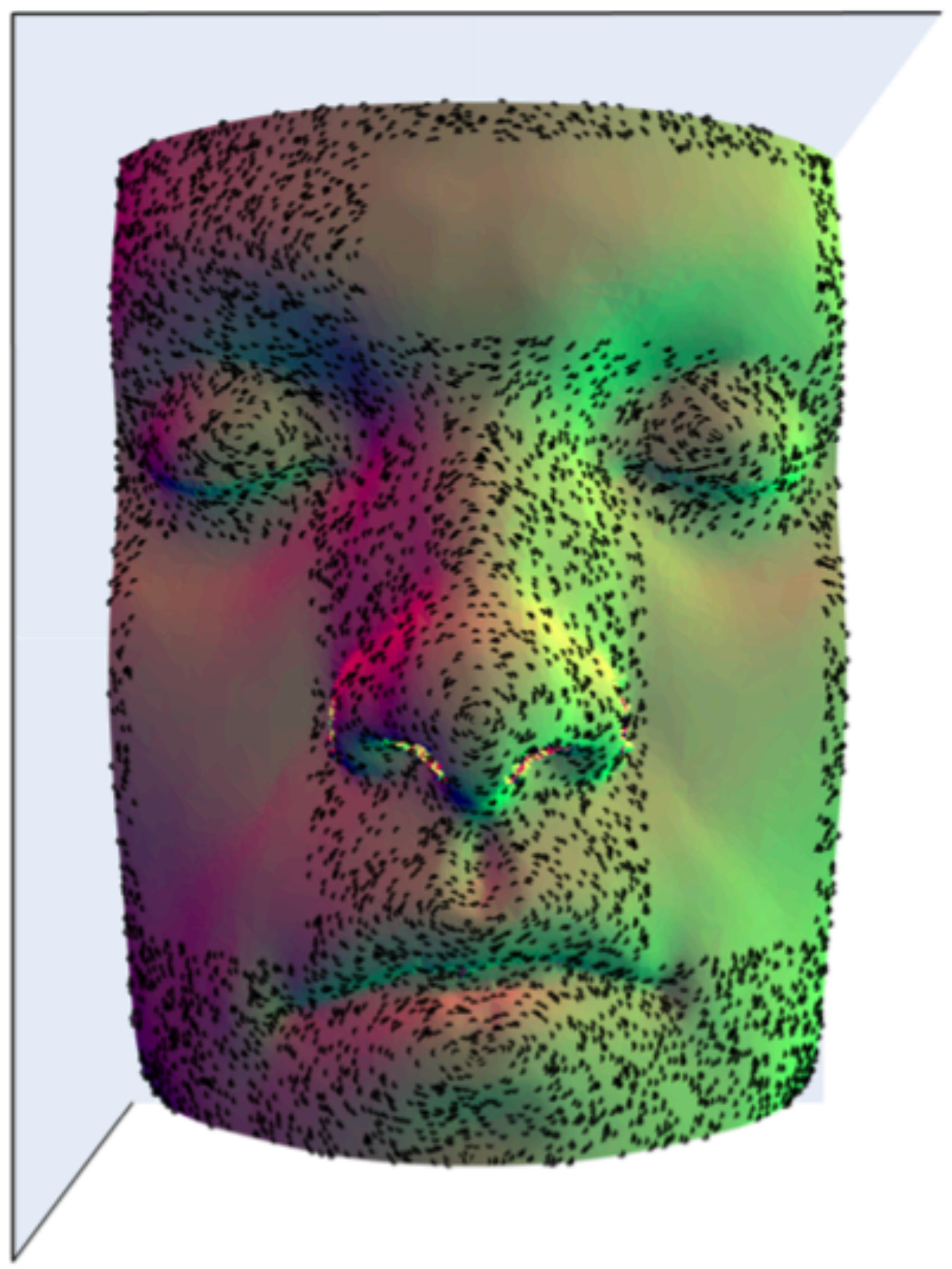
(III) Rank-1 structure of the Hessian

Example: HTV As a Complexity Measure

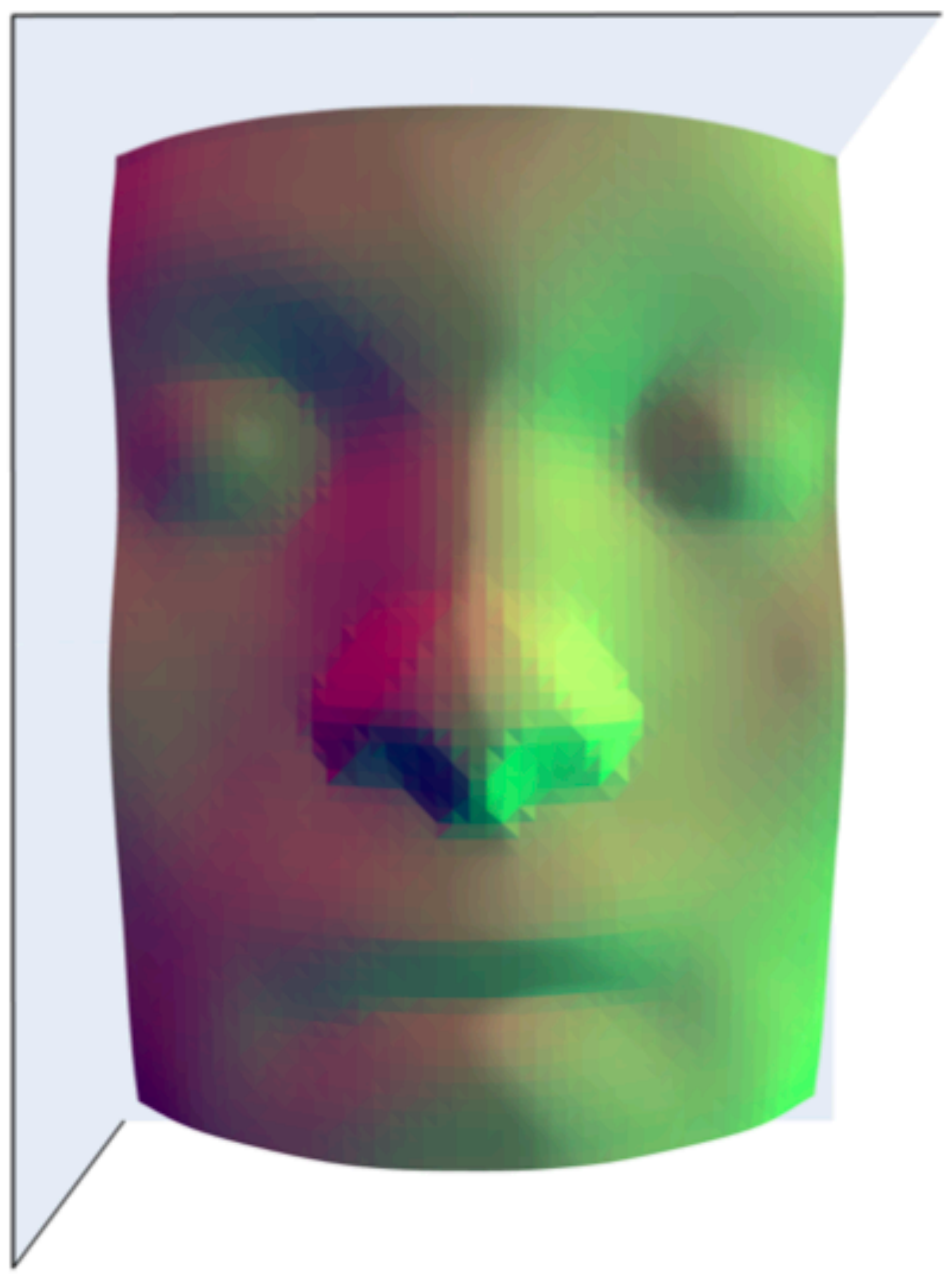
- In dimension $d = 1$: $\mathrm{HTV}_p(f) = \mathrm{TV}^{(2)}(f)$



Example: HTV As a Complexity Measure

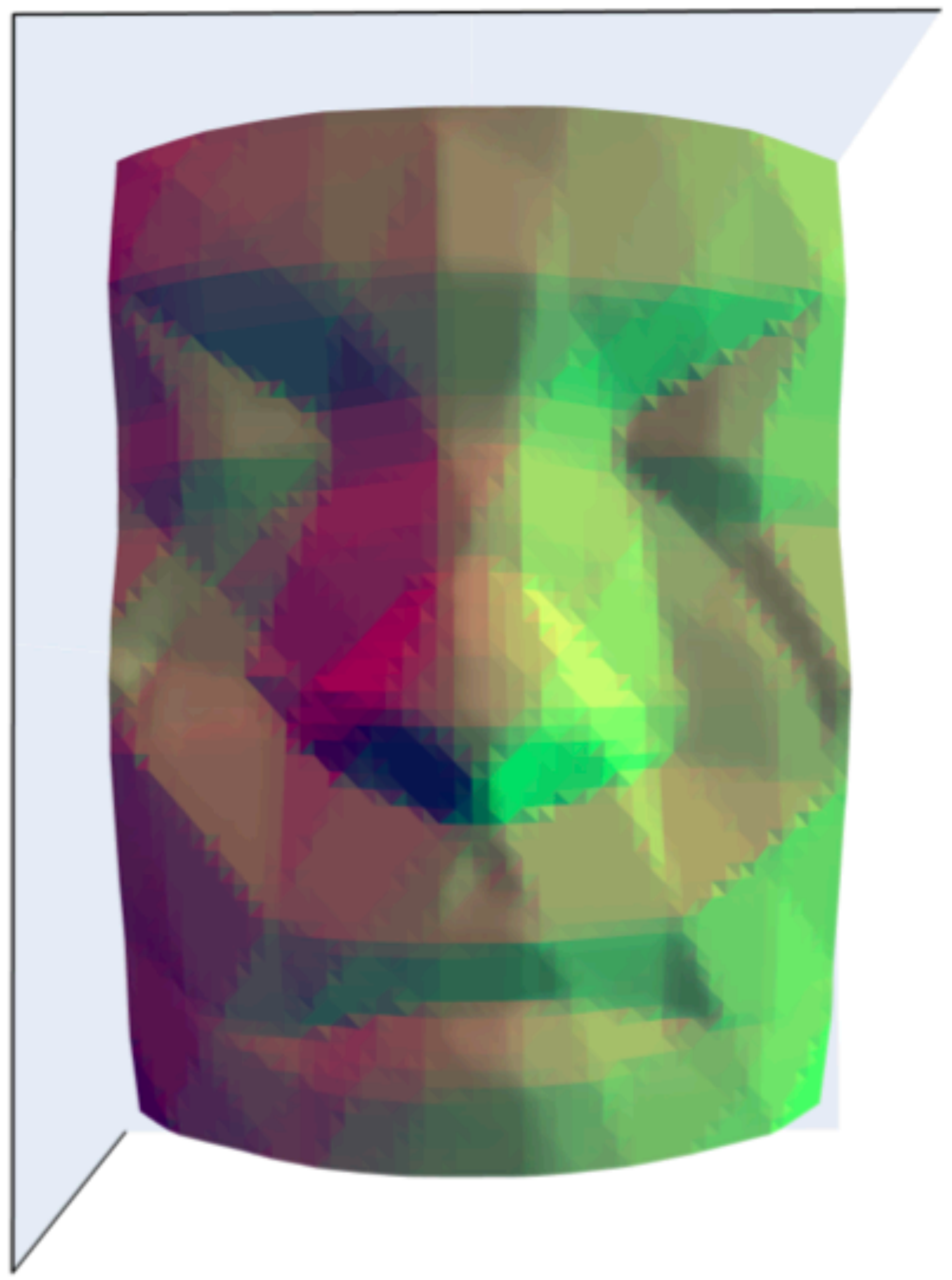


Target function
+
M=5000 training data

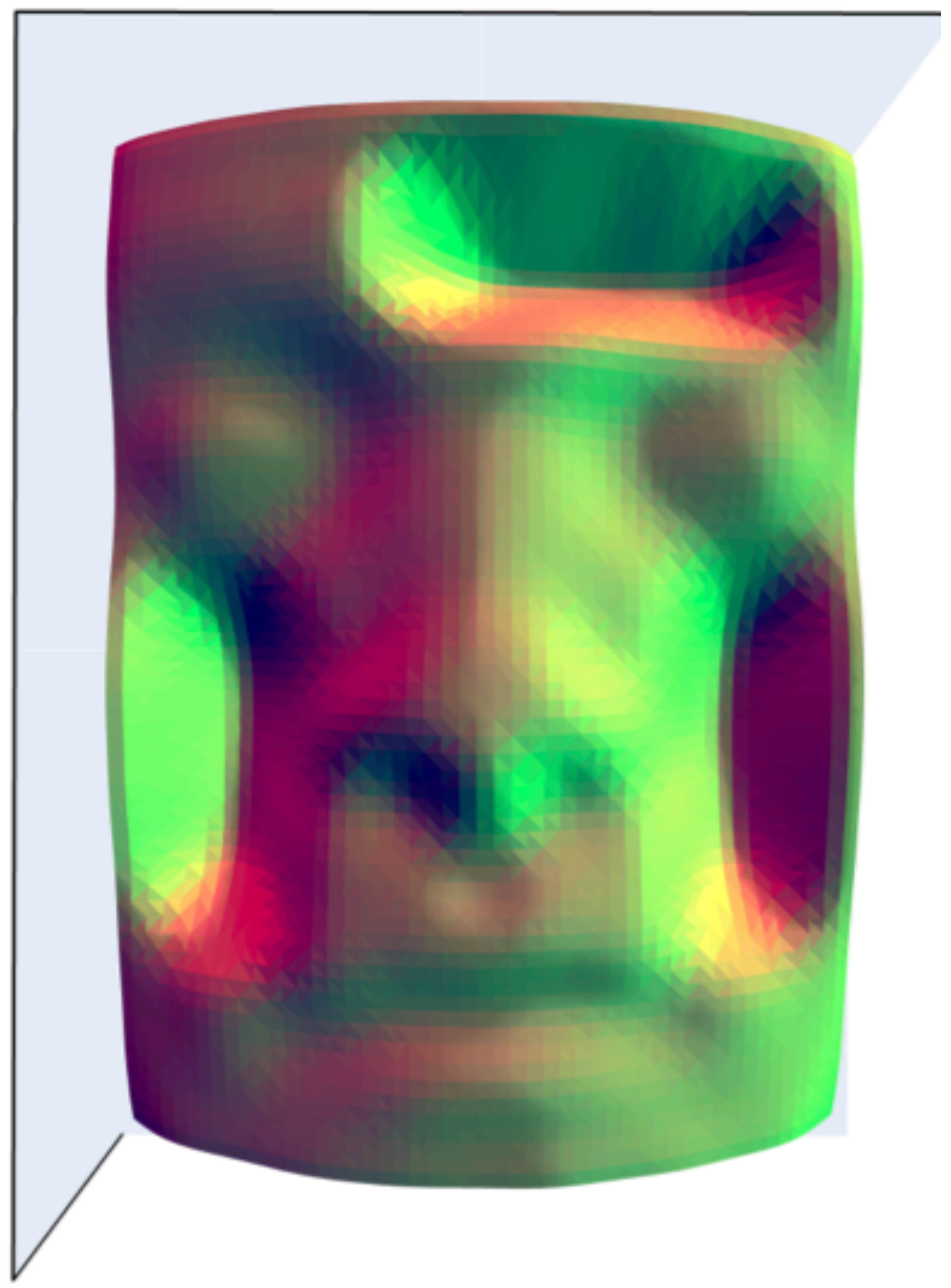


HTV Min

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

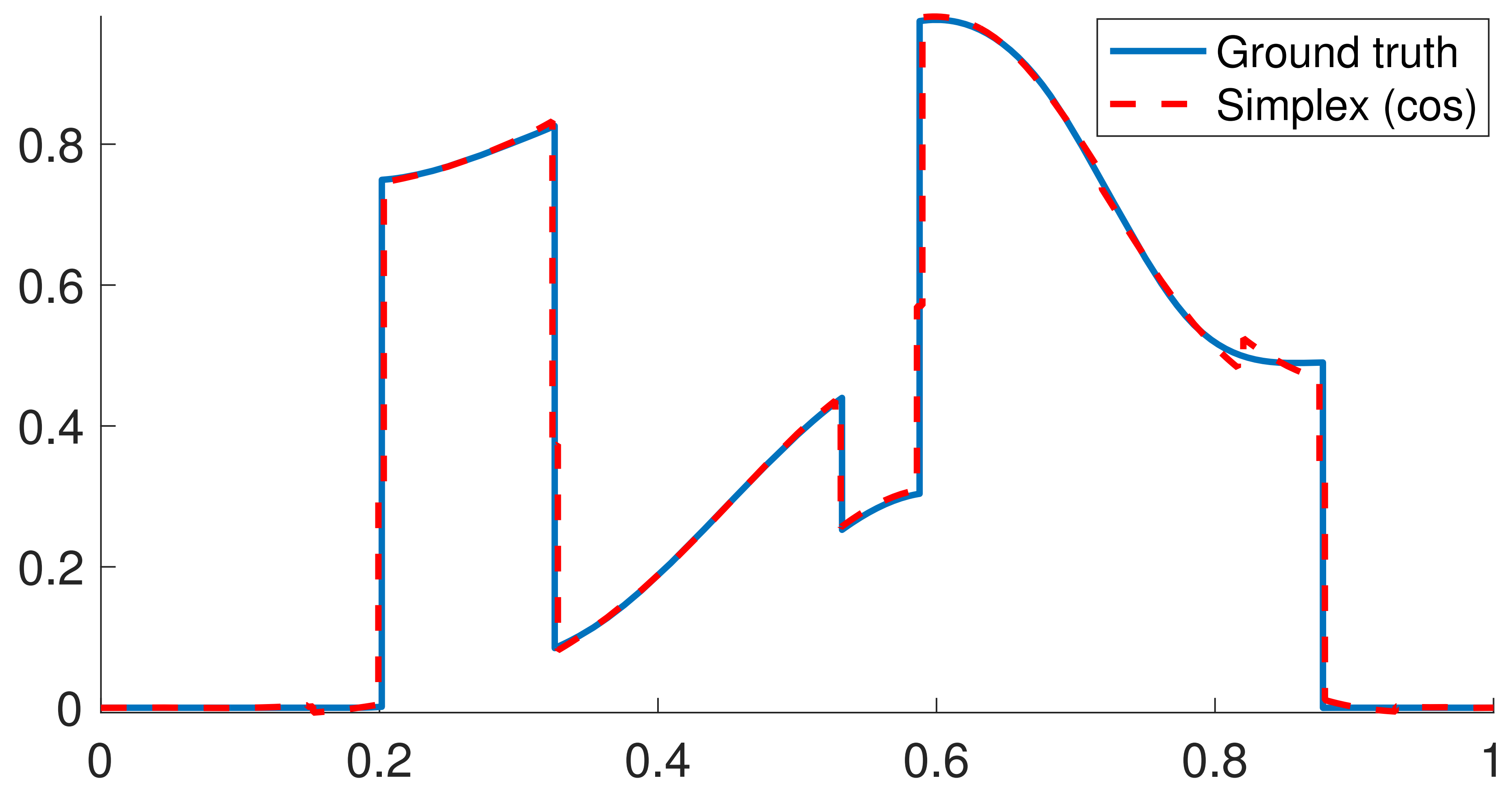
Part III: Multicomponent Inverse Problems

- Multicomponent model: $s = s_1 + s_2$

    1. Both components are sparse, albeit in different domains

    2. One component is sparse, the other one is smooth

    3. Application: 2D curve fitting



- Relevant publications

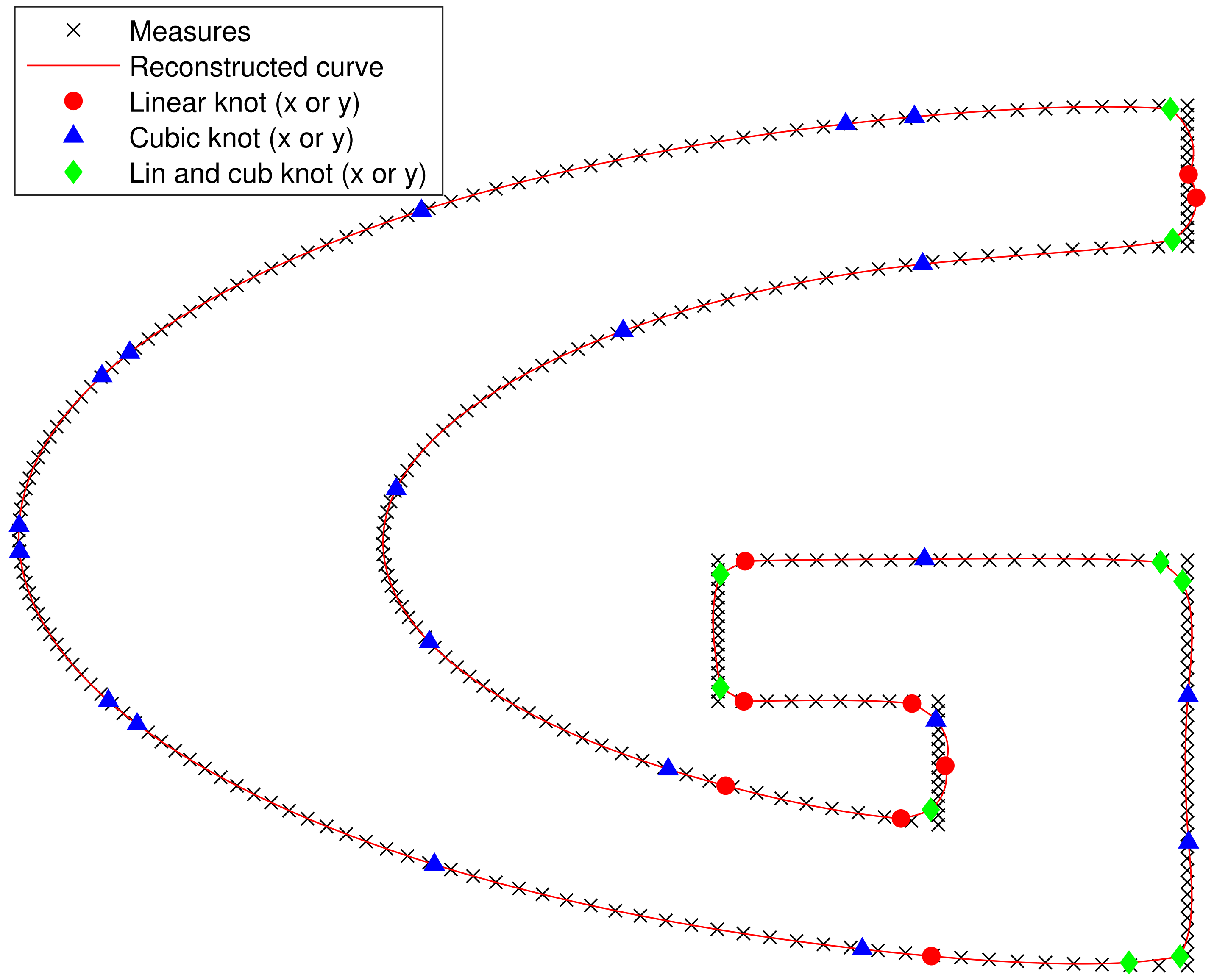
  - T. Debarre, **S. Aziznejad**, M. Unser, "Hybrid-spline dictionaries for continuous-domain inverse problems," *IEEE Transactions on Signal Processing, 2019*.

  - T. Debarre, **S. Aziznejad**, M. Unser, "Continuous-domain formulation of inverse problems for composite sparse-plus-smooth signals," *IEEE Open Journal of Signal Processing, 2021*.

  - I. Lloréns Jover, T. Debarre, **S. Aziznejad**, M. Unser, "Coupled splines for sparse curve fitting," *ArXiv, 2021*.

Part III: Multicomponent Inverse Problems

- Multicomponent model: $s = s_1 + s_2$



3. Application: 2D curve fitting

- Relevant publications

  - I. Lloréns Jover, T. Debarre, **S. Aziznejad**, M. Unser, "Coupled splines for sparse curve fitting," *ArXiv, 2021*.

2D Curve Fitting

- Goal: Find $\mathbf{r}(t) = (x(t), y(t))$ that best fits $\mathbf{p}[m] = (p_x[m], p_y[m])$

- Our formulation: curve fitting as an inverse problem

- Regularization functional: $\mathcal{R}(\mathrm{L}\{\mathbf{r}\})$
  - $\mathrm{L} = \mathrm{D}^N$, $N \geq 2$
  - $\mathcal{R}$: A novel rotation-invariant mixed-norm

Definition [Lloréns Jover et al. '21]

Let $p \in [1, +\infty]$ and $q = p/(p-1)$. The $\mathrm{TV} - \ell_p$ mixed-norm of $\mathbf{w} = (w_1, w_2) \in \mathcal{S}'(\mathbb{T})^2$ is defined as

$$\|\mathbf{w}\|_{\mathrm{TV}-\ell_p} = \sup\left\{\langle \mathbf{w}, \boldsymbol{\varphi}\rangle : \boldsymbol{\varphi} \in \mathcal{S}(\mathbb{T})^2, \|\boldsymbol{\varphi}(x)\|_q \leq 1 \forall \boldsymbol{x} \in \mathbb{T}\right\}.$$

Proposition [Lloréns Jover et al. '21]

The $\mathrm{TV} - \ell_p$ mixed-norm is rotation invariant, if and only if $p = 2$.

- $\mathcal{R} = \|\cdot\|_{\mathrm{TV}-\ell_2}$

2D Curve Fitting

Theorem [Lloréns Jover et al. '21]

1. For any curve $\mathbf{f} = (f_1, f_2)$ with absolutely integrable components $f_i \in L_1(\mathbb{T}_M)$, $i = 1, 2$, we have that

$$\|[f_1 \quad f_2]\|_{\mathrm{TV}-\ell_p} = \int_0^M \|\mathbf{f}(t)\|_p \mathrm{d}t.$$

2. Let $\mathbf{w} = (w_1, w_2)$ be a vector-valued distribution of the form $\mathbf{w} = \sum_{k=1}^{K} \mathbf{a}[k] Ш_M(\cdot - t_k)$ with $\mathbf{a}[k] \in \mathbb{R}^2, k = 0, \dots, K-1$. Then, we have that

$$\|[w_1 \quad w_2]\|_{\mathrm{TV}-\ell_p} = \sum_{k=0}^{K-1} \|\mathbf{a}[k]\|_p.$$

Theorem [Lloréns Jover et al. '21]
There is a hybrid-spline solution with $K \leq 2M + 2$ knots for the minimization

$$\min_{\substack{\mathbf{r}_i \in \mathcal{X}_{\mathrm{L}_i}(\mathbb{T}_M) \\ \mathbf{r}_1(0)=\mathbf{0}}} \sum_{m=0}^{M-1} \|\mathbf{r}_1(t)|_{t=m} + \mathbf{r}_2(t)|_{t=m} - \mathbf{p}[m]\|_2^2 + \lambda_1 \|\mathrm{L}_1\{\mathbf{r}_1\}\|_{\mathrm{TV}-\ell_2} + \lambda_2 \|\mathrm{L}_2\{\mathbf{r}_2\}\|_{\mathrm{TV}-\ell_2}.$$

Sketch of proof

Item 1 and 2:

(I) LHS $\leq$ RHS

- Hölder's inequality

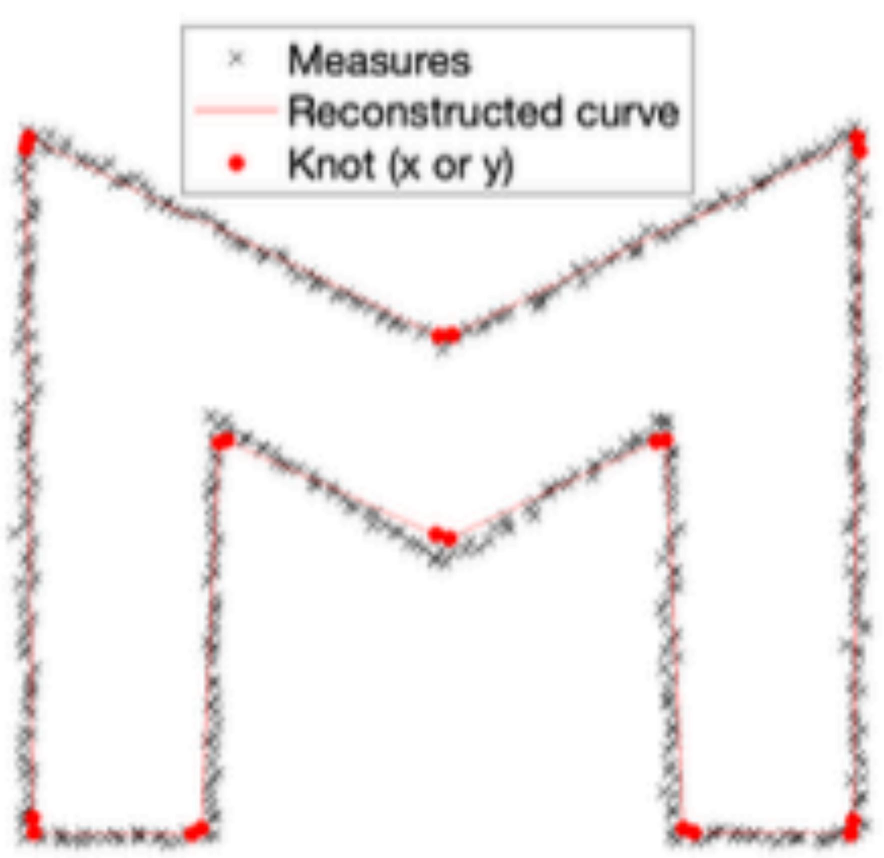
(II) $\forall \epsilon > 0$: LHS $\geq$ RHS $- \epsilon$

- Lusin's theorem
- Duality mapping of ℓ_p norms

Sketch of proof

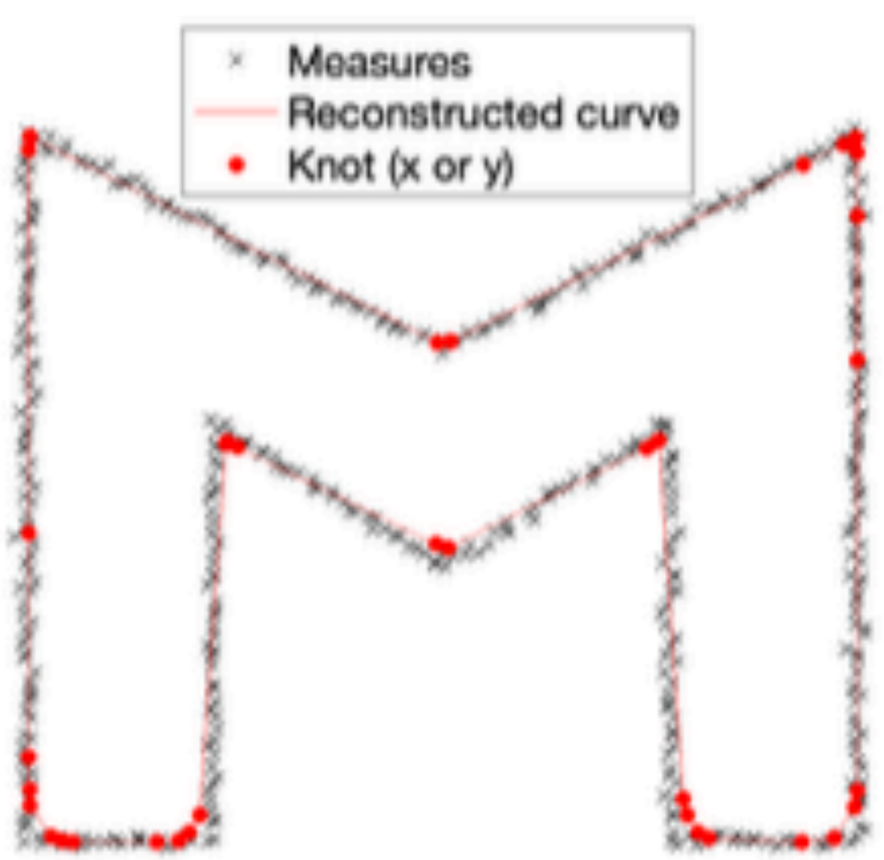
1. Existence
   - Direct-product
   - seminorm minimization
2. Form of the solution
   - Extreme points of the RI-TV ball

Example

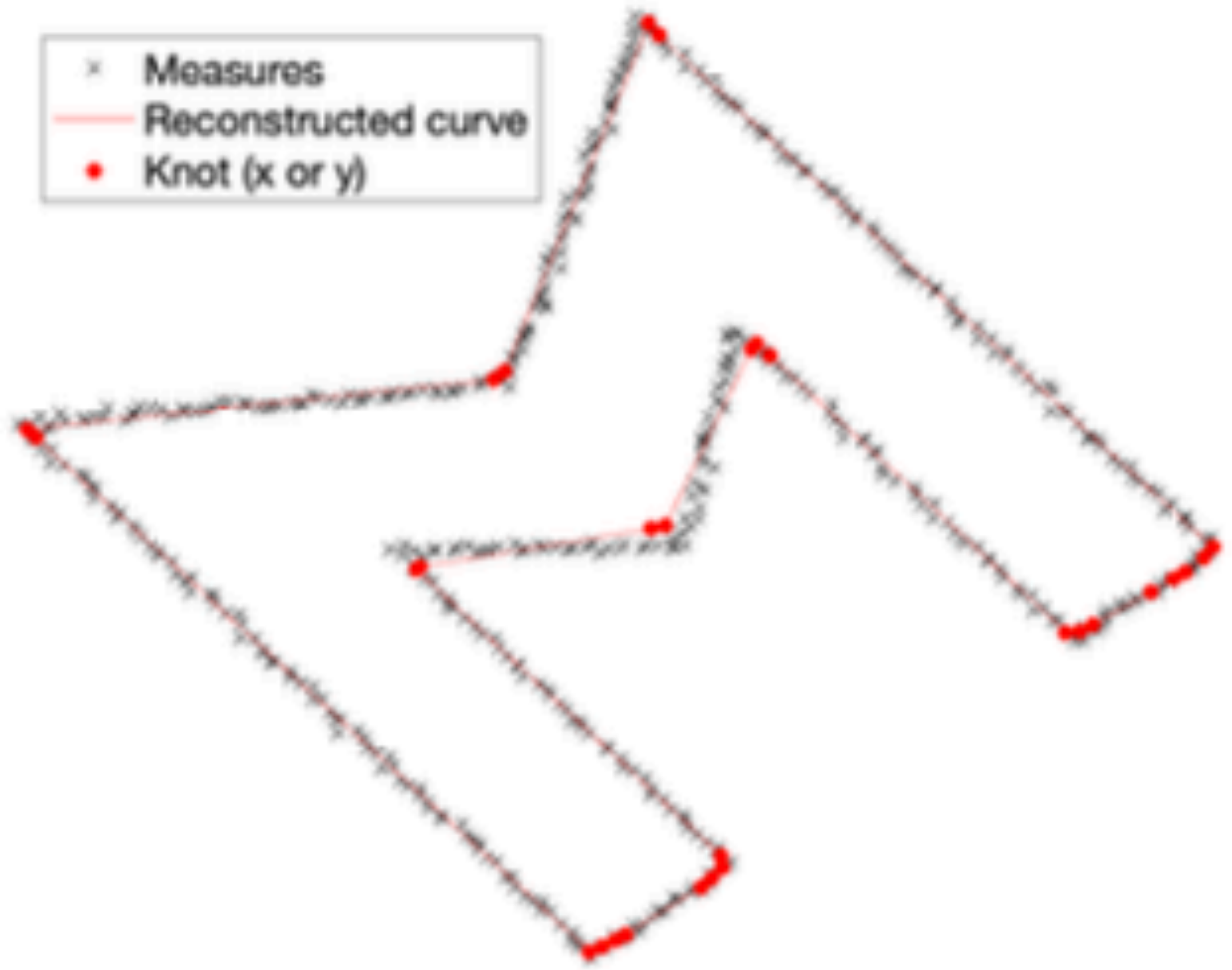


(a) RI-TV regularization, $\theta = 0°$,
$K = 20$, $\lambda = 700$,
QFE $= 12.09$.

(b) RI-TV regularization, $\theta = 40°$,
$K = 20$, $\lambda = 700$,
QFE $= 12.09$.

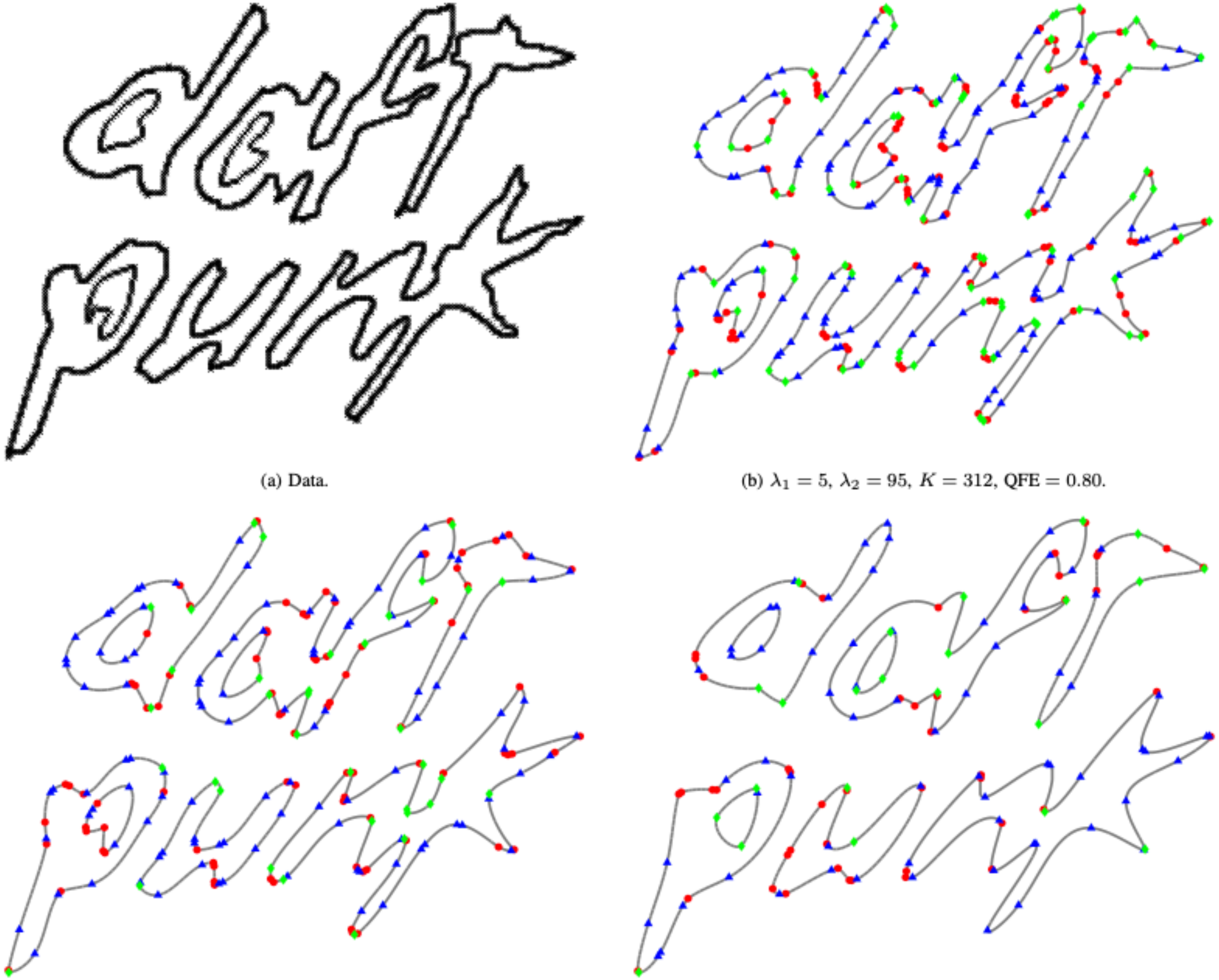


(c) (TV-ℓ_1) regularization, $\theta = 0°$,
$K = 37$, $\lambda = 482.13$,
QFE $= 12.09$.



(d) (TV-ℓ_1) regularization, $\theta = 40°$,
$K = 29$, $\lambda = 500.93$,
QFE $= 12.09$.

Example



(a) Data.

(b) $\lambda_1 = 5,\ \lambda_2 = 95,\ K = 312,\ \mathrm{QFE} = 0.80.$

(c) $\lambda_1 = 20,\ \lambda_2 = 980,\ K = 229,\ \mathrm{QFE} = 1.11.$

(d) $\lambda_1 = 100,\ \lambda_2 = 9900,\ K = 139,\ \mathrm{QFE} = 2.82.$

Conclusion

[O] Convex optimization problems over Banach spaces

O1. Direct-product search spaces

O2. Seminorm regularization

[L] Supervised learning with sparsity prior

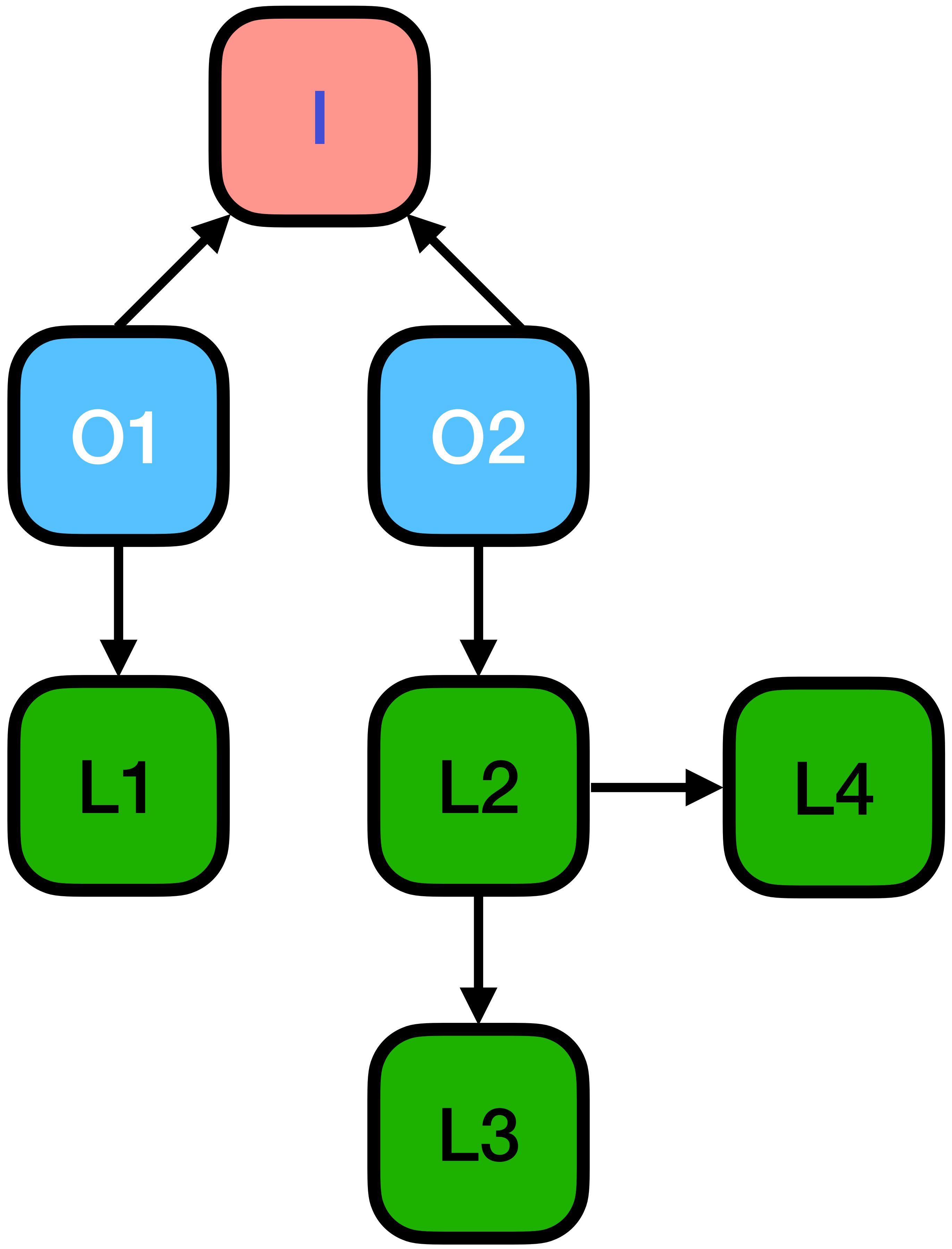
L1. Sparse multikernel regression

L2. Univariate learning with sparsity and Lipschitz constraint

L3. Learning activation functions of deep neural networks

L4. Learning multivariate CPWL functions with HTV regularization

[I] Multicomponent inverse problems



Many thanks!