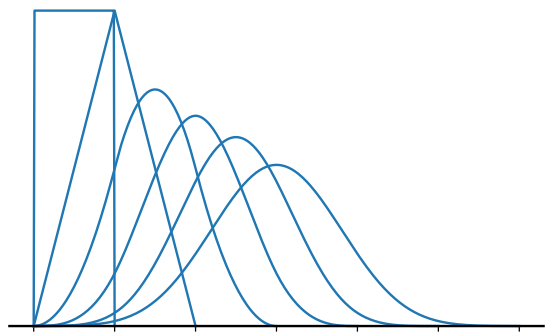
# On BV Spaces, Splines, and Neural Networks

Rahul Parhi
Department of Electrical and Computer Engineering
University of Wisconsin–Madison

(joint work with Robert Nowak)

Analysis Seminar

November 16th, 2021



# Outline


1. Introduction
   - Reconstructing functions from measurements
   - Variational methods for function reconstruction
   - Splines and variational methods in $\mathrm{BV}$ spaces
2. Neural networks and variational methods
   - New $\mathrm{BV}$ spaces related to neural networks
3. What are these new spaces?

# UW–Madison is (was?) the mecca of splines

- **Isaac Schoenberg** (1903–1990) invented the spline in the 1940s.

  $\Longrightarrow$ Was at UW–Madison from 1966–1990.
- **Carl de Boor** wrote many influential papers and books about approximation theory and numerical algorithms with splines.

  $\Longrightarrow$ Was at UW–Madison from 1972–2003.
- **Grace Wahba** wrote many influential papers about smoothing noisy data with splines, making splines popular in statistics.

  $\Longrightarrow$ Was at UW–Madison from 1967–2018.
- **Amos Ron** has written many influential papers about approximation theory with splines.

  $\Longrightarrow$ Has been at UW–Madison since 1988.

**Remark**

People seem to have forgotten about splines...

# Reconstructing functions from measurements

- A fundamental problem in science and engineering is to **reconstruct** a function $f : \mathbb{R}^d \to \mathbb{R}$ from **measurements**.

$$y_n = \langle h_n, f \rangle + \varepsilon_n,\ n = 1, \ldots, N$$

- $\mathrm{H}\{f\} = (\langle h_1, f \rangle, \ldots, \langle h_N, f \rangle) \in \mathbb{R}^N$ symbolizes the **linear** measurement process.
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N) \in \mathbb{R}^N$ are perturbation or noise terms, typically zero-mean random variables.
- $\boldsymbol{y} = (y_1, \ldots, y_N) \in \mathbb{R}^N$ denotes the (possibly noisy) data.

**Remark**

This is an **ill-posed inverse problem**.

# How do we solve this problem?

- Solving this problem requires choosing/designing a model for your functions.
  $\Longrightarrow$ Assume that $f \in \mathcal{X}'$, where $\mathcal{X}'$ is some Banach space.
  $\Longrightarrow$ It will be useful to think of $\mathcal{X}'$ as a dual space.
- To remedy the ill-posed nature of the reconstruction problem, a **minimum-energy** requirement is often imposed to **regularize** the solution

$$\min_{f \in \mathcal{X}'} \|f\|_{\mathcal{X}'} \quad \text{s.t.} \quad \langle h_n, f \rangle = y_n,\ n = 1, \ldots, N.$$

or, if the data is noisy,

$$\min_{f \in \mathcal{X}'} \sum_{n=1}^{N} |y_n - \langle h_n, f \rangle|^2 + \lambda \|f\|_{\mathcal{X}'}^p,$$

where $\lambda > 0$ and $1 \leq p < \infty$.
$\Longrightarrow$ $\|\cdot\|_{\mathcal{X}'}$ is a norm or seminorm that defines $\mathcal{X}'$ and is often referred to as a **regularizer**.

# Representer theorems

$$\min_{f \in \mathcal{X}'} \|f\|_{\mathcal{X}'} \quad \text{s.t.} \quad \langle h_n, f \rangle = y_n,\ n = 1, \ldots, N.$$

$$\min_{f \in \mathcal{X}'} \sum_{n=1}^{N} |y_n - \langle h_n, f \rangle|^2 + \lambda \|f\|_{\mathcal{X}'}^p,$$

- Representer theorems provide a **parametric representation**–ideally, a finite linear expansion in terms of some "basis" functions or **atoms**—that span the solution set.

  $\Longrightarrow$ Representer theorems provide a way to recast **infinite-dimensional** optimization problems as **finite-dimensional** optimization problems, providing a first step to designing a **numerical method** to solve the infinite-dimensional problem.

- This variational formulation for function reconstruction captures many classical methods such as **smoothing splines**, **locally adaptive splines,** and **wavelet methods**.

# (Cubic) smoothing splines

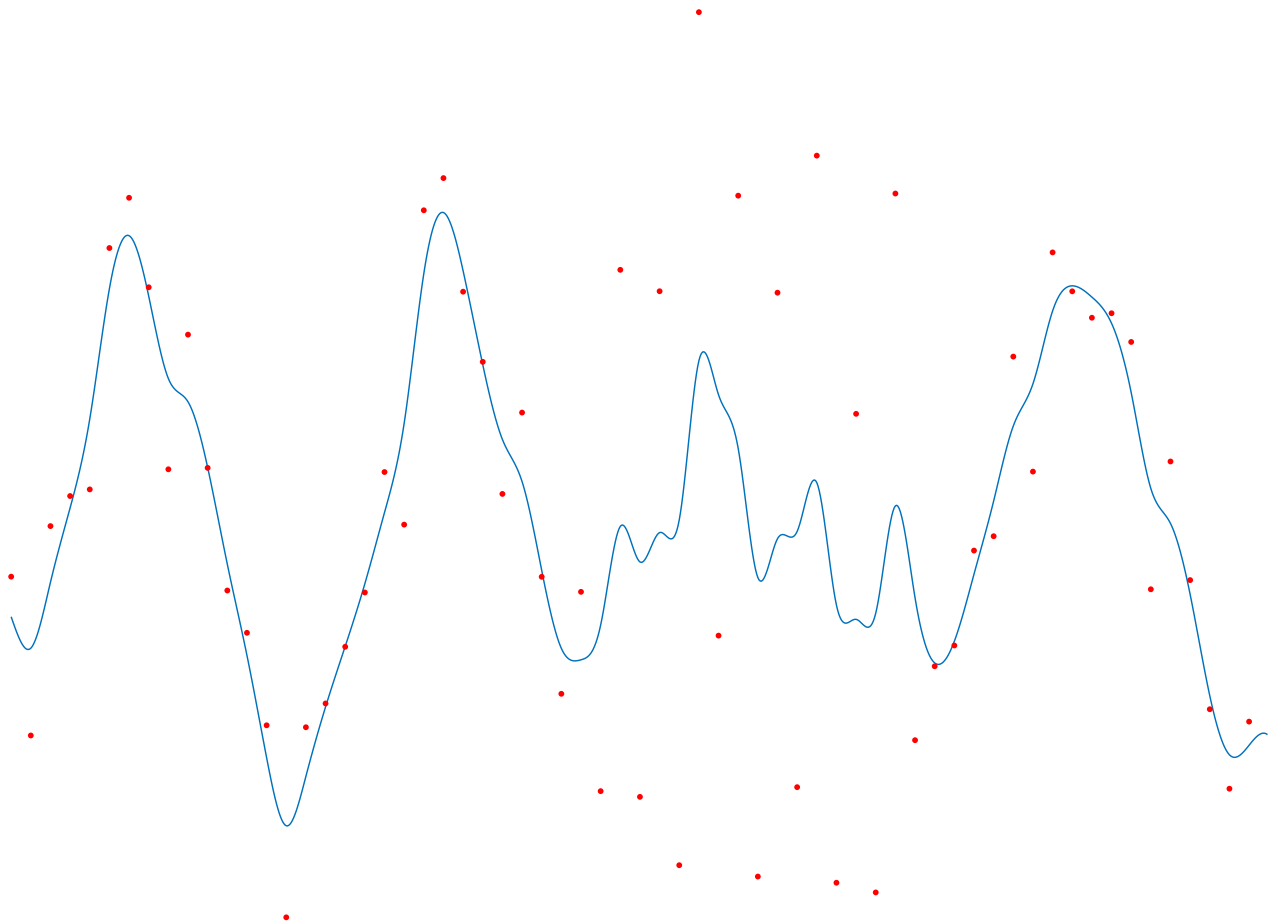
$$\min_{f \in \dot{H}^2(\mathbb{R})} \sum_{n=1}^{N} |y_n - f(x_n)|^2 + \lambda \|\mathrm{D}^2 f\|_{L^2}^2$$

- Measurements are point evaluations $h_n = \delta(\cdot - x_n)$ for some $\{x_n\}_{n=1}^N \subset \mathbb{R}$.
- The solution to this problem is **unique** and a **cubic spline**

$$x \mapsto \sum_{n=1}^{N} v_k \, (x - x_n)_+^3 + c_1 x + c_0$$

- The number of **knots** is equal to the number of data.
- The **atoms** of the solution are $(\cdot - x_n)_+^3$.

# (Cubic) smoothing splines



# (Linear) locally adaptive splines

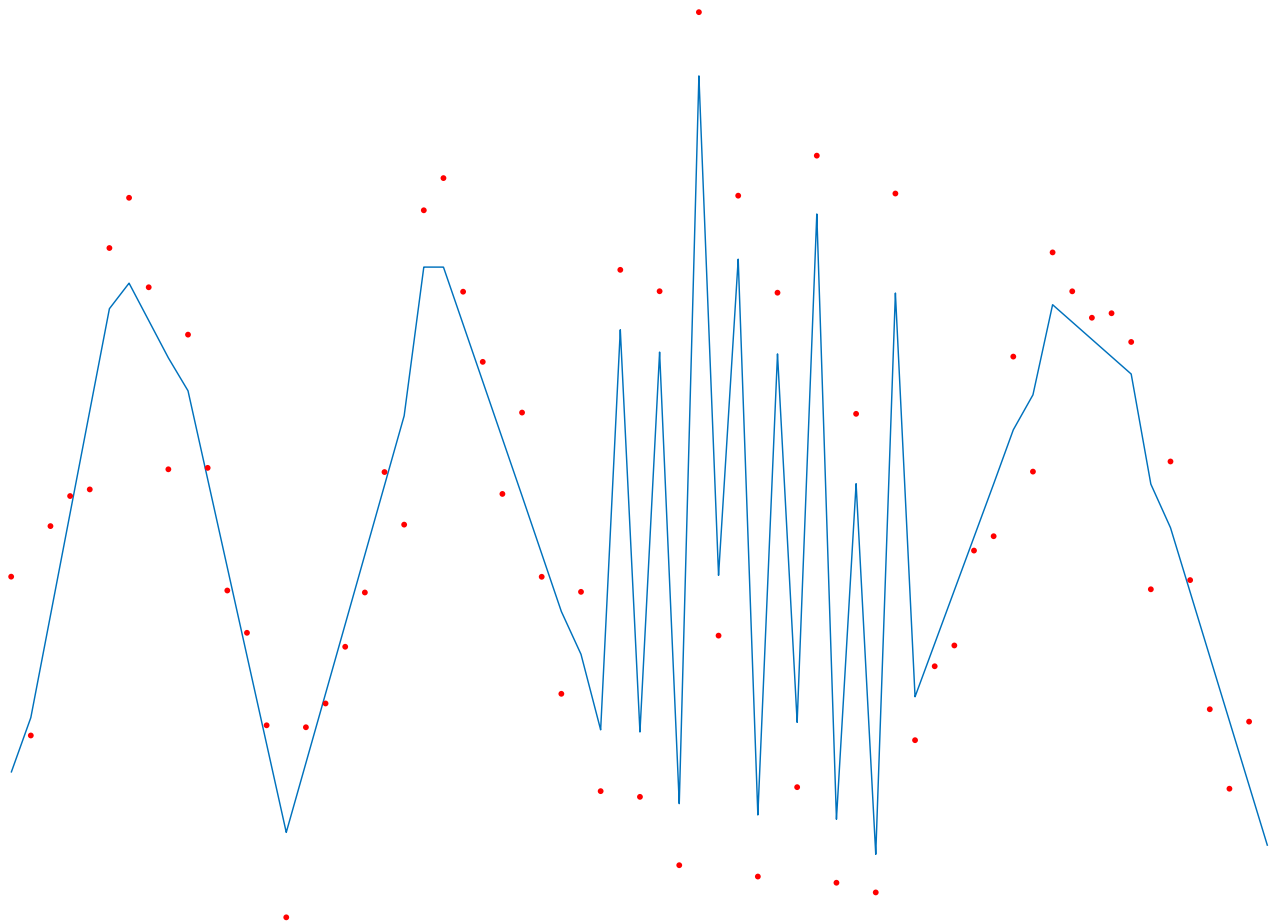
$$\min_{f:\mathrm{D}\, f\in \mathrm{BV}(\mathbb{R})} \sum_{n=1}^{N} |y_n - f(x_n)|^2 + \lambda\, \mathrm{TV}(\mathrm{D}\, f)$$

- Measurements are point evaluations $h_n = \delta(\cdot - x_n)$ for some $\{x_n\}_{n=1}^N \subset \mathbb{R}$.
- There **exists** a solution to this problem that is a **linear spline**

$$x \mapsto \sum_{k=1}^{K} v_k\, (x - t_k)_+ + c_1 x + c_0$$

- The number of **knots** is $K \leq N$.
- The **atoms** of the solution are $(\cdot - t_k)_+$.

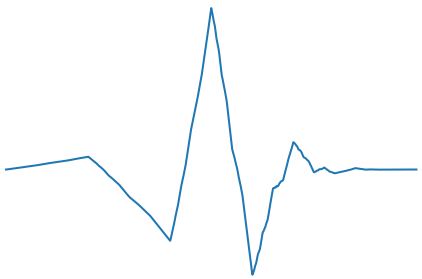
# (Linear) locally adaptive splines



# Wavelet methods

- Wavelet atoms are translates and dilates of a **mother wavelet** function $\psi$

$$\left\{\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)\right\}_{j,k\in\mathbb{Z}}$$



- The Daubechies wavelets provide orthobases for $L^2(\mathbb{R})$.
- $\|f\|_{B^s_{p,q}} \asymp \left(\sum_j 2^{jsq}\|\Delta_j f\|^q_{L^p}\right)^{1/q}$
- Wavelet characterizations of $B^s_{p,q}(\mathbb{R})$ essentially replace $\Delta_j$ with

$$P_j f = \sum_{k\in\mathbb{Z}} \langle f, \psi_{j,k}\rangle \psi_{j,k}$$

# (Db3) wavelet methods

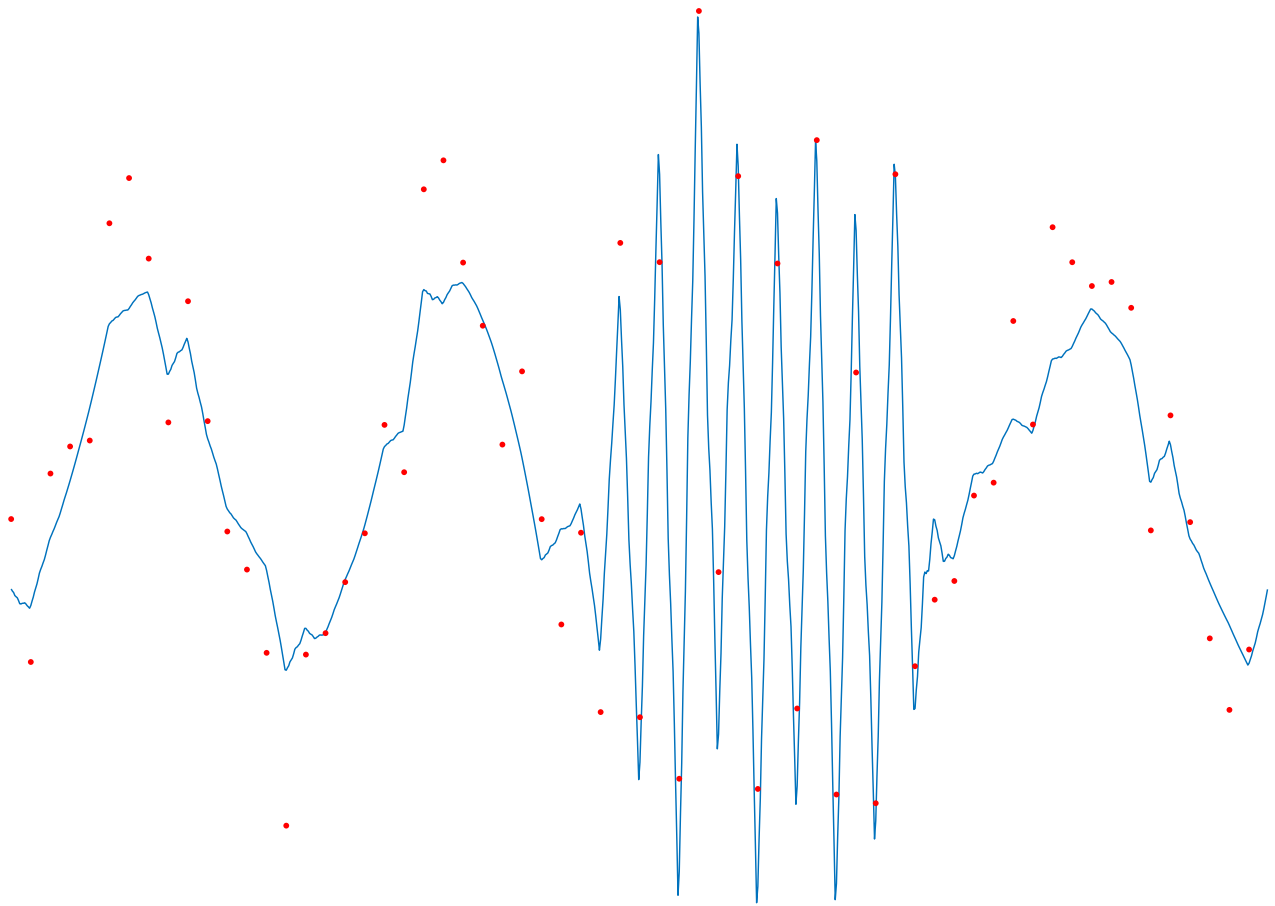
$$\min_{f \in B^2_{1,1}(\mathbb{R})} \sum_{n=1}^{N} |y_n - f(x_n)|^2 + \lambda \|f\|_{B^2_{1,1}}$$

- $\|\cdot\|_{B^2_{1,1}}$ is the db3 wavelet dependent norm.
- Measurements are point evaluations $h_n = \delta(\cdot - x_n)$ for some $\{x_n\}_{n=1}^N \subset \mathbb{R}$.
- There **exists** a solution to this problem that can be written in terms of a finite number of wavelet functions

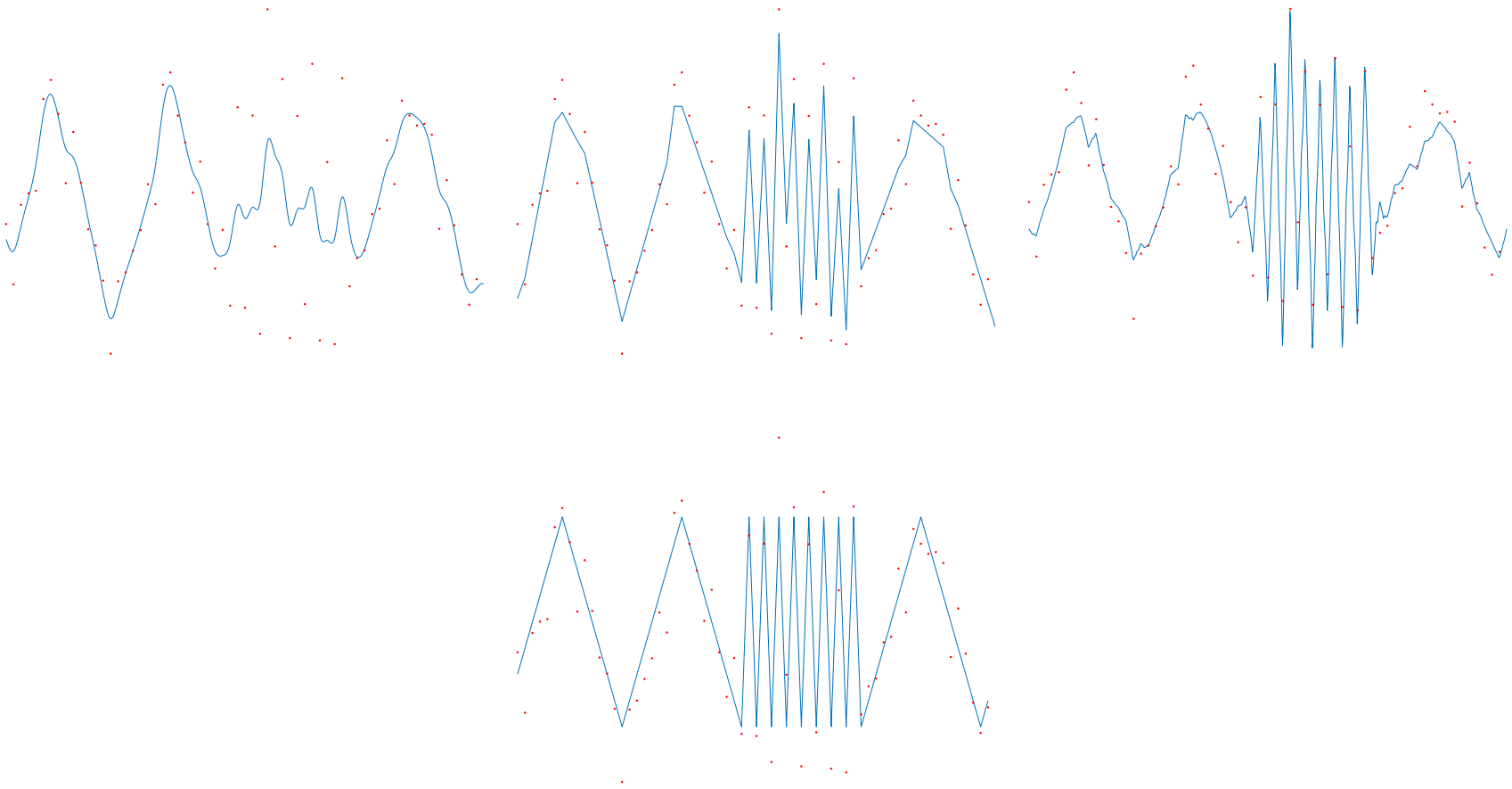
$$x \mapsto \sum_{(j,k) \in \mathcal{I}} c_{j,k} \psi_{j,k}(x)$$

- $|\mathcal{I}|$ depends on $\lambda$.
- The **atoms** of the solution are $\psi_{j,k}$.

# (Db3) wavelet methods



# Comparing these methods



- The data-generating function is not in $\dot{H}^2(\mathbb{R})$!

  $\Longrightarrow$ Choosing the right function space is **crucial** in getting accurate reconstructions.

# Variational formulation for function reconstruction

- **Neural networks** are outperforming and replacing classical methods in many reconstruction tasks.
- Unlike classical methods, neural networks are not well-understood mathematically.

  $\Longrightarrow$ Classical methods assume regularity of the underlying function and design a procedure to optimally reconstruct functions with that regularity.

  $\Longrightarrow$ Neural network methods do not aim assume any kind of regularity a priori, but outperform classical methods in practice.
- It turns out that this variational formulation also captures neural networks. (P. and Nowak 2020, 2021a,b,c)

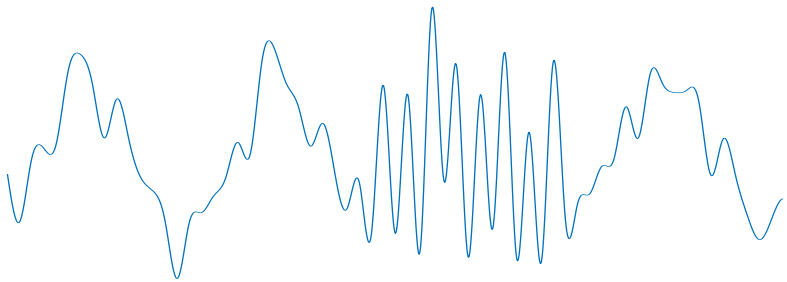
$$\min_{f \in \mathcal{X}'} \|f\|_{\mathcal{X}'} \quad \text{s.t.} \quad \langle h_n, f \rangle = y_n,\ n = 1, \ldots, N.$$

$\Longrightarrow$ $\mathcal{X}'$ is a **new**, not a classically studied function space.

$\Longrightarrow$ $\mathcal{X}'$ is a new kind of $\mathrm{BV}$ space that captures the regularity properties that neural networks are intrinsically optimal for.

# Spatial inhomogenity and sparsity

- Many real-world objects (signals, images, functions, etc.) are **spatially inhomogeneous** or exhibit **sparsity**.
  $\Longrightarrow$ Spatial inhomogeneity is modeled via Besov spaces with $p < 2$.
  $\Longrightarrow$ Sparsity is modeled via Besov spaces with $p = 1$ or with $\mathrm{BV}$-type spaces.



- e.g., $\mathrm{BV}$ is a common model for natural images.

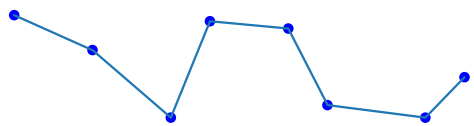
$$B^1_{1,1}(\Omega) \subset \mathrm{BV}(\Omega) \subset B^1_{1,\infty}(\Omega)$$

- These kinds of spaces are interesting from an analysis perspective since they are **non-reflexive**.
- The spaces related to neural networks are non-reflexive.
  $\Longrightarrow$ Before understanding neural networks, we first need to understand (locally adaptive) splines.

# What are splines?

- A **polynomial spline** of degree $m-1$ is a piecewise polynomial function that is $C^{m-2}$.
- The locations of the discontinuities of the $(m-1)$th derivative are referred to as the **knots** of the spline.
- A function $f$ is a polynomial spline of degree $m-1$ if

$$\mathrm{D}^m f = \sum_{k=1}^{K} v_k \, \delta(\cdot - t_k),$$



where $m$ is the **order** of the spline and $\{t_k\}_{k=1}^K$ are the **knots**.

- Therefore, every spline can be written as

$$x \mapsto \sum_{k=1}^{K} v_k \, \rho_m(x - t_k) + \sum_{n=0}^{m-1} c_n \, x^n,$$

where $\rho_m(x) = x_+^{m-1}/(m-1)!$ is a **Green's function** of $\mathrm{D}^m$.
$\Longrightarrow$ $\rho_m$ is the $m$th-order **truncated power function**.

# Notation

- For $f : \mathbb{R} \to \mathbb{R}$, let $\mathrm{TV}(f) = \sup_{x_1 < \cdots < x_N} \sum_{n=0}^{N} |f(x_n) - f(x_{n-1})|$.
- $\mathcal{M}(\mathbb{R}) \subset \mathscr{S}'(\mathbb{R})$ be the space of finite Radon measures on $\mathbb{R}$.

$$\|u\|_{\mathcal{M}} = \sup_{\substack{\varphi \in C_0(\mathbb{R}) \\ \|\varphi\|_\infty = 1}} \langle u, \varphi \rangle$$

- $\|\cdot\|_{\mathcal{M}}$ is the total variation norm in the sense of measures.
- $\mathrm{TV}(f) = \|\mathrm{D}\, f\|_{\mathcal{M}}$
- Let $\mathrm{TV}^m(f) = \mathrm{TV}(\mathrm{D}^{m-1} f) = \|\mathrm{D}^m f\|_{\mathcal{M}}$ denote the $m$th-order total variation of $f$.
- Let

$$\mathrm{BV}^m(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} : \ \mathrm{TV}^m(f) < \infty\}$$

denote the space of functions of $m$th-order $\mathrm{BV}$ on $\mathbb{R}$.

$\Longrightarrow$ This is a **non-reflexive Banach space**.

# Representer theorem for locally adaptive splines

## Theorem (Fisher and Jerome 1975)

Consider the variational problem

$$\mathcal{V} := \underset{f \in \mathrm{BV}^m(\mathbb{R})}{\arg\min} \ \|\mathrm{D}^m f\|_{\mathcal{M}} \quad \text{s.t.} \quad \langle h_n, f \rangle = y_n,\ n = 1, \ldots, N,$$

where $h_n : \mathrm{BV}^m(\mathbb{R}) \to \mathbb{R}$ is weak$^*$ continuous, i.e., $h_n$ lies in the predual of $\mathrm{BV}^m(\mathbb{R})$. Then, the solution set $\mathcal{V}$ is nonempty, convex, and weak$^*$ compact. Moreover, the extreme points of $\mathcal{V}$ are $m$th-order splines of the form

$$s(x) = \sum_{k=1}^{K} v_k \, \rho_m(x - t_k) + \sum_{n=0}^{m-1} c_n \, x^n, \quad K \leq N.$$

- These solutions are **sparse**: $\|\mathrm{D}^m s\|_{\mathcal{M}} = \|\boldsymbol{v}\|_1$ and $K \leq N$.
- The knot locations $\{t_k\}_{k=1}^{K}$ are **adaptive**.

# Proof sketch

**Step 1:** Understanding the Banach structure of $\mathrm{BV}^m(\mathbb{R})$.

- $\mathrm{BV}^m(\mathbb{R})$ is defined by the seminorm $f \mapsto \|\mathrm{D}^m f\|_{\mathcal{M}}$
- The null space of this seminorm is the space of polynomials of degree $< m$, denoted $\mathcal{P}_{m-1}(\mathbb{R})$.
- We can equip $\mathcal{P}_{m-1}(\mathbb{R})$ with a **biorthogonal system** so that every $q \in \mathcal{P}_{m-1}(\mathbb{R})$ admits a unique representation

$$q = \sum_{k=1}^{m} \langle \phi_k, q \rangle p_k,$$

where $\{p_k\}_{k=1}^m$ is a basis for $\mathcal{P}_{m-1}(\mathbb{R})$ and $\phi_k \in \mathcal{P}_{m-1}(\mathbb{R})'$ such that $\langle \phi_k, p_n \rangle = \delta_{kn}$, e.g., choose $p_k(x) = \frac{x^{k-1}}{(k-1)!}$ and $\phi_k(x) = (-1)^{k-1}\delta^{(k-1)}(x)$.

# Proof sketch (cont.)

- Consider the operator $\mathrm{D}_{\boldsymbol{\phi}}^{-m}$ whose Schwartz kernel is given by

$$\begin{aligned} g_{\boldsymbol{\phi}}(x, y) &= \rho_m(x - y) - \sum_{k=1}^{m} \langle \phi_k, \rho_m(\cdot - y) \rangle p_k(x) \\ &= \rho_m(x - y) - \mathrm{proj}_{\mathcal{P}_{m-1}(\mathbb{R})}\{\rho_m(\cdot - y)\}(x) \end{aligned}$$

- For all $\mu \in \mathcal{M}(\mathbb{R})$, this operator satisfies

$$\begin{aligned} \mathrm{D}^m \, \mathrm{D}_{\boldsymbol{\phi}}^{-m} \, \mu &= \mu \\ \boldsymbol{\phi}(\mathrm{D}_{\boldsymbol{\phi}}^{-m} \, \mu) &= \mathbf{0}, \end{aligned}$$

where $\boldsymbol{\phi}(f) = (\langle \phi_1, f \rangle, \ldots, \langle \phi_m, f \rangle)$.

# Proof sketch (cont.)

- Define

  $$\mathrm{BV}^m_{\boldsymbol{\phi}}(\mathbb{R}) = \{f \in \mathrm{BV}^m(\mathbb{R}) : \ \boldsymbol{\phi}(f) = \mathbf{0}\} \cong \mathrm{BV}^m(\mathbb{R})/\mathcal{P}_{m-1}(\mathbb{R}).$$

  Therefore, $\mathrm{BV}^m(\mathbb{R}) = \mathrm{BV}^m_{\boldsymbol{\phi}}(\mathbb{R}) \oplus \mathcal{P}_{m-1}(\mathbb{R})$.
- Since we factored out the null space, $\mathrm{BV}^m_{\boldsymbol{\phi}}(\mathbb{R})$ is a Banach space when equipped with the norm $f \mapsto \|\mathrm{D}^m f\|_{\mathcal{M}}$.
- Consider the mapping $\mathrm{D}^m : \mathrm{BV}^m_{\boldsymbol{\phi}}(\mathbb{R}) \to \mathcal{M}(\mathbb{R})$ that maps $f \mapsto \mu = \mathrm{D}^m f$. Since $\|\mathrm{D}^m f\|_{\mathcal{M}} = \|\mu\|_{\mathcal{M}}$, this map is norm preserving and it is invertible with inverse given by $\mathrm{D}^{-m}_{\boldsymbol{\phi}} : \mathcal{M}(\mathbb{R}) \to \mathrm{BV}^m_{\boldsymbol{\phi}}(\mathbb{R})$.

  $\Longrightarrow$ i.e., $\mathrm{BV}^m_{\boldsymbol{\phi}}(\mathbb{R}) \cong \mathcal{M}(\mathbb{R})$.

# Proof sketch (cont.)

- Every $f \in \mathrm{BV}^m(\mathbb{R})$ admits the direct-sum decomposition

$$f = \mathrm{D}_{\boldsymbol{\phi}}^{-m} \mu + q,$$

where $\mu = \mathrm{D}^m f \in \mathcal{M}(\mathbb{R})$ and $q = \mathrm{proj}_{\mathcal{P}_{m-1}(\mathbb{R})} f \in \mathcal{P}_{m-1}(\mathbb{R})$.

- When equipped with the norm

$$\|f\|_{\mathrm{BV}^m(\mathbb{R})} = \|\mathrm{D}^m f\|_{\mathcal{M}} + \|\boldsymbol{\phi}(f)\|_2,$$

$\mathrm{BV}^m(\mathbb{R})$ is a Banach space, in particular, a non-reflexive Banach space (since $\mathcal{M}(\mathbb{R})$ is a non-reflexive Banach space).

$\Longrightarrow$ e.g., $\|f\|_{\mathrm{BV}(\mathbb{R})} = \mathrm{TV}(f) + |f(0)|$.

# Proof sketch (cont.)

**Step 2:** Transform the variational problem over $\mathrm{BV}^m(\mathbb{R})$ into a variational problem over $\mathcal{M}(\mathbb{R})$.

- Recall the variational problem

$$\min_{f \in \mathrm{BV}^m(\mathbb{R})} \|\mathrm{D}^m f\|_{\mathcal{M}} \quad \text{s.t.} \quad \mathrm{H}\{f\} = \boldsymbol{y} \in \mathbb{R}^N.$$

- From the direct-sum decomposition $f = \mathrm{D}_{\boldsymbol{\phi}}^{-m} \mu + q$, this problem is equivalent to

$$\min_{\mu \in \mathcal{M}(\mathbb{R})} \|\mu\|_{\mathcal{M}} \quad \text{s.t.} \quad \mathrm{H}\,\mathrm{D}_{\boldsymbol{\phi}}^{-m} \mu = \boldsymbol{y} - \mathrm{H}\, q \in \mathbb{R}^N.$$

# Proof sketch (cont.)

- Classical problem known as **Radon measure recovery**:

$$\mathcal{V} := \operatorname*{arg\,min}_{\mu \in \mathcal{M}(\mathbb{R})} \|\mu\|_{\mathcal{M}} \quad \text{s.t.} \quad \mathcal{A}\mu = \boldsymbol{z},$$

where $\mathcal{A} : \mathcal{M}(\mathbb{R}) \to \mathbb{R}^N$ is weak$^*$ continuous and linear.

$\Longrightarrow$ It is well-known (Fisher and Jerome 1975; Zuhovickiĭ 1948) that the solution set $\mathcal{V}$ is nonempty, convex, and weak$^*$ compact, and that the extreme points take the form

$$\sum_{k=1}^{K} v_k \, \delta(\cdot - t_k),$$

where $t_k \in \mathbb{R}$, $k = 1, \ldots, K$, and $K \leq N$.

$\Longrightarrow$ weak$^*$ continuity of the measurement operator plays a crucial role in proving that solutions exist.

# Proof sketch (cont.)

- Therefore, the solution set to

$$\min_{f \in \mathrm{BV}^m(\mathbb{R})} \|\mathrm{D}^m f\|_{\mathcal{M}} \quad \text{s.t.} \quad \mathrm{H}\{f\} = \boldsymbol{y} \in \mathbb{R}^N.$$

nonempty, convex, and weak$^*$ compact. Moreover, the extreme points are $m$th-order splines of the form

$$s(x) = \sum_{k=1}^{K} v_k \, \rho_m(x - t_k) + \sum_{n=0}^{m-1} c_n \, x^n,$$

where $K \leq N$.

$\Longrightarrow$ Plug in the solution to the Radon measure recovery problem into the direct-sum decomposition. $\square$

# What's going on?

$$\min_{f \in \mathrm{BV}^m(\mathbb{R})} \|\mathrm{D}^m f\|_{\mathcal{M}} \quad \text{s.t.} \quad \mathrm{H}\{f\} = \boldsymbol{y} \in \mathbb{R}^N.$$

- The extreme points of the solution set are written as a superposition of extreme points of the the unit ball associated to $f \mapsto \|\mathrm{D}^m f\|_{\mathcal{M}}$ (in the quotient space $\mathrm{BV}^m(\mathbb{R})/\mathcal{P}_{m-1}(\mathbb{R})$).

$\Longrightarrow$ We showed that $\mathrm{BV}^m_{\boldsymbol{\phi}}(\mathbb{R}) \cong \mathrm{BV}^m(\mathbb{R})/\mathcal{P}_{m-1}(\mathbb{R})$ is isometrically isomorphic to $\mathcal{M}(\mathbb{R})$.

$\Longrightarrow$ It is well-known that the extreme points of the unit ball of $\mathcal{M}(\mathbb{R})$ take the form $\delta(\cdot - t_0)$, $t_0 \in \mathbb{R}$.

$\Longrightarrow$ Therefore, the extreme points of $\mathrm{BV}^m_{\boldsymbol{\phi}}(\mathbb{R})$ take the form

$$\begin{aligned} \mathrm{D}^{-m}_{\boldsymbol{\phi}}\{\delta(\cdot - t_0)\} &= g_{\boldsymbol{\phi}}(\cdot, t_0) \\ &= \rho_m(\cdot - t_0) - \mathrm{proj}_{\mathcal{P}_{m-1}(\mathbb{R})}\{\rho_m(\cdot - t_0)\} \\ &\in [\rho_m(\cdot - t_0)] \end{aligned}$$

$\Longrightarrow$ This unit ball has an interesting **geometry**.

# A general representer theorem

## Theorem

- Let $(\mathcal{X}, \mathcal{X}')$ be a dual pair of Banach spaces.
- Let $h_1, \ldots, h_N \in \mathcal{X}$ be a set of measurement functionals.
- Let $\mathrm{H} : \mathcal{X}' \to \mathbb{R}^N : f \mapsto (\langle h_1, f \rangle, \ldots, \langle h_N, f \rangle)$ be a weak$^*$ continuous linear measurement operator.

Then, for any fixed $\boldsymbol{y} \in \mathbb{R}^N$, the solution set to

$$\operatorname*{arg\,min}_{f \in \mathcal{X}'} \|f\|_{\mathcal{X}'} \quad \text{s.t.} \quad \mathrm{H}\{f\} = \boldsymbol{y}$$

is nonempty, convex, and weak$^*$ compact, and its extreme points take the form

$$\sum_{k=1}^{K} v_k u_k,$$

where $K \leq N$ and $u_k \in \mathrm{Ext}\{f \in \mathcal{X}' : \ \|f\|_{\mathcal{X}'} \leq 1\}$.

# A general representer theorem

Variants of this result have been proven by a number of authors:

- Claire Boyer et al. (2019). "On representer theorems and convex regularization". In: **SIAM Journal on Optimization** 29.2, pp. 1260–1281
- Kristian Bredies and Marcello Carioni (2020). "Sparsity of solutions for variational inverse problems with finite-dimensional data". In: **Calculus of Variations and Partial Differential Equations** 59.1, pp. 1–26
- Michael Unser (2021). "A unifying representer theorem for inverse problems and machine learning". In: **Foundations of Computational Mathematics** 21.4, pp. 941–960

# What is a neural network?

- Compositions of affine mappings and nonlinear mappings.
- A neural network $f : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{L+1}}$ can be written as

$$\begin{cases} \boldsymbol{x}^{(0)} := \boldsymbol{x}, \\ \boldsymbol{x}^{(\ell)} := \boldsymbol{\rho}(\mathbf{A}^{(\ell)}\boldsymbol{x}^{(\ell-1)} - \boldsymbol{b}^{(\ell)}),\ \ell = 1, \dots, L, \\ f(\boldsymbol{x}) = \mathbf{A}^{(L+1)}\boldsymbol{x}^{(L)} - \boldsymbol{b}^{(L+1)}, \end{cases}$$

$\Longrightarrow$ $L \in \mathbb{N}$ is the number of **hidden layers**.
$\Longrightarrow$ $\mathbf{A}^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ are the **weights** of the neural network.
$\Longrightarrow$ $\boldsymbol{b}^{(\ell)} \in \mathbb{R}^{d_\ell}$ are the **biases** of the neural network.
$\Longrightarrow$ $\rho : \mathbb{R} \to \mathbb{R}$ is the **activation function**; $\boldsymbol{\rho}$ applies $\rho$ entrywise.

**Remark**

When $L = 1$, the neural network is **shallow** and when $L > 1$, the neural network is **deep**.

# What is a neural network?

- Today we will focus on **shallow** neural networks with **scalar** outputs.
- These are functions $f : \mathbb{R}^d \to \mathbb{R}$ that can be written as

$$f(\boldsymbol{x}) = \boldsymbol{v}^\mathsf{T} \boldsymbol{\rho}(\mathbf{W}\boldsymbol{x} - \boldsymbol{b}) = \sum_{k=1}^{K} v_k \, \rho(\boldsymbol{w}_k^\mathsf{T} \boldsymbol{x} - b_k),$$

  where $v_k \in \mathbb{R}$, $\boldsymbol{w}_k \in \mathbb{R}^d$, and $b_k \in \mathbb{R}$.
- A common choice for $\rho$ is the truncated linear function $\rho(x) = x_+$, which we previously defined as $\rho_2$.
- The atoms of a shallow neural network are functions of the form $\rho(\boldsymbol{w}_k^\mathsf{T}(\cdot) - b_k)$ and are often referred to as **neurons**.

**Goal**

Prove a representer theorem for neural networks.

# Observations

- Recall the variational problem for locally adaptive splines

$$\min_{f \in \mathrm{BV}^m(\mathbb{R})} \|\mathrm{D}^m f\|_{\mathcal{M}} \quad \text{s.t.} \quad \langle h_n, f \rangle = y_n,\ n = 1, \ldots, N.$$

- The solution set is completely characterized by $m$th-order splines of the form

$$x \mapsto \sum_{k=1}^{K} v_k\, \rho_m(x - t_k) + \sum_{n=0}^{m-1} c_n\, x^n, \quad \rho_m(x) = \frac{x_+^{m-1}}{(m-1)!}.$$

- The operator $\mathrm{D}^m$ is a **sparsifying transform** to the atom $\rho_m(\cdot - t_k)$ since it **extracts** the **parameter** of the atom

$$\mathrm{D}^m\, \rho_m(\cdot - t_k) = \delta(\cdot - t_k).$$

- In the proof of the representer theorem we were able to construct $\mathrm{D}_{\boldsymbol{\phi}}^{-m}$, a **bounded, right inverse** of $\mathrm{D}^m$.

# Observations

- The atoms of a neural network take the form

$$\boldsymbol{x} \mapsto \rho(\boldsymbol{w}_0^\mathsf{T}\boldsymbol{x} - b_0), \quad \rho(x) = x_+$$

- Univariate neural networks with $\rho(x) = x_+$ are **linear splines**

$$\implies \mathrm{D}^2\left\{\sum_{k=1}^K v_k\,\rho(w_k(\cdot) - b_k)\right\} = \sum_{k=1}^K v_k|w_k|\delta(\cdot - b_k/w_k).$$

- The parameters of an atom are $\boldsymbol{w}_0 \in \mathbb{R}^d$ and $b_0 \in \mathbb{R}$.

$\implies$ WLOG, suppose $\boldsymbol{w}_0 \in \mathbb{S}^{d-1} = \left\{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_2 = 1\right\}$.

$$\implies \rho(\boldsymbol{w}_0^\mathsf{T}\boldsymbol{x} - b_0) = \|\boldsymbol{w}_0\|_2\,\rho\left(\left[\frac{\boldsymbol{w}_0}{\|\boldsymbol{w}_0\|_2}\right]^\mathsf{T}\boldsymbol{x} - \frac{b_0}{\|\boldsymbol{w}_0\|_2}\right)$$

# Question

## Question

Does there exist a sparsifying transform $\mathrm{R}$ such that

$$\mathrm{R}\,\rho(\boldsymbol{w}_0^{\mathsf{T}}(\cdot) - b_0) = \delta(\cdot - (\boldsymbol{w}_0, b_0)),$$

where $(\boldsymbol{w}_0, b_0) \in \mathbb{S}^{d-1} \times \mathbb{R}$?

## Answer

Yes, and it involves **second-order differentiation** and the **Radon transform**.

# The Radon transform

- For a sufficiently nice function $f : \mathbb{R}^d \to \mathbb{R}$, the **Radon transform** is given by

$$\mathscr{R}\{f\}(\zeta) = \int_\zeta f,$$

  $\implies$ $\zeta \in \mathbb{P}^d$ is a $(d-1)$-dimensional hyperplane in $\mathbb{R}^d$.
- Every hyperplane can be defined as the solution set to the equation $\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x} = t$, where $(\boldsymbol{\gamma}, t) \in \mathbb{S}^{d-1} \times \mathbb{R}$.
  $\implies$ $(\boldsymbol{\gamma}, t)$ and $(-\boldsymbol{\gamma}, -t)$ are associated to the same hyperplane.
- We can identify $\mathscr{R}\{f\}$ with an even function on $\mathbb{S}^{d-1} \times \mathbb{R}$ given by

$$\mathscr{R}\{f\}(\boldsymbol{\gamma}, t) = \int_{\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x}=t} f(\boldsymbol{x})\,\mathrm{d}s(\boldsymbol{x})$$

- The Radon transform as an integral operator whose Schwartz kernel is given by the distribution

$$k(\boldsymbol{x}, (\boldsymbol{\gamma}, t)) = \delta(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x} - t)$$

# Inverting the dual Radon transform

- For a sufficiently nice function $\Phi : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$, the **dual Radon transform** is given by

$$\mathscr{R}^*\{\Phi\}(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}} \Phi(\boldsymbol{\gamma}, \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x}) \, \mathrm{d}\sigma(\boldsymbol{\gamma})$$

- Given sufficiently nice $\Phi : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ such that $\Phi(\boldsymbol{\gamma}, t) = \Phi(-\boldsymbol{\gamma}, -t)$,

$$2(2\pi)^{d-1}\Phi = \Lambda^{d-1}\,\mathscr{R}\,\mathscr{R}^*\,\Phi = \mathscr{R}\,\mathscr{R}^*\,\Lambda^{d-1}\Phi,$$

where $\Lambda^{d-1} = (-\partial_t^2)^{\frac{d-1}{2}}$.

$\Longrightarrow$ In CT imaging parlance, $\Lambda^{d-1}$ is known as the **backprojection filter** (or ramp filter).

# The sparsifying transform

**Lemma (Ongie et al., 2020, P. and Nowak, 2021)**

Consider $r_{(\boldsymbol{w}_0,b_0)} = \rho(\boldsymbol{w}_0^\mathsf{T}(\cdot) - b_0)$, where $\rho(x) = x_+$. Then,

$$c_d \, \Lambda^{d-1} \, \mathscr{R} \, \Delta \, r_{(\boldsymbol{w}_0,b_0)} = \frac{\delta(\cdot - (\boldsymbol{w}_0, b_0)) + \delta(\cdot + (\boldsymbol{w}_0, b_0))}{2}$$

where $c_d \coloneqq \frac{1}{2(2\pi)^{d-1}}$.

# Proof

- Notice that

$$\Delta\, r_{(\boldsymbol{w}_0, b_0)} = \Delta\, \rho(\boldsymbol{w}_0^{\mathsf{T}}(\cdot) - b_0) = \delta(\boldsymbol{w}_0^{\mathsf{T}}(\cdot) - b_0),$$

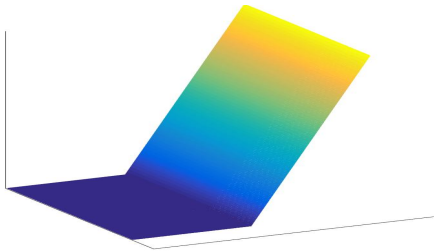
  which is the Schwartz kernel of the Radon transform.
- Next, consider the even test function $\psi$.
- Therefore,

$$\begin{aligned} c_d \langle \Lambda^{d-1}\, \mathscr{R}\, \Delta\, r_{(\boldsymbol{w}_0, b_0)}, \psi \rangle &= c_d \langle \Delta r_{(\boldsymbol{w}_0, b_0)}, \mathscr{R}^*\, \Lambda^{d-1} \psi \rangle \\ &= (c_d\, \mathscr{R}\, \mathscr{R}^*\, \Lambda^{d-1} \psi)(\boldsymbol{w}_0, b_0) \\ &= \psi(\boldsymbol{w}_0, b_0) \end{aligned}$$

# What's going on?

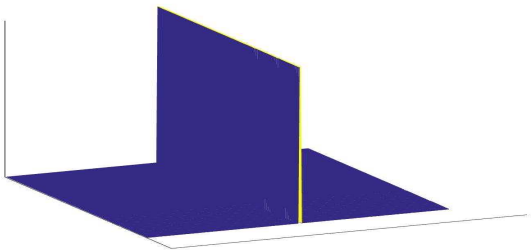
- Neural network atom

$$\Longrightarrow \quad \rho(\boldsymbol{w}_0^{\mathsf{T}}(\cdot) - b_0),\ (\boldsymbol{w}_0, b_0) \in \mathbb{S}^{d-1} \times \mathbb{R}$$



- **Laplacian** of atom

$$\Longrightarrow \quad \Delta\big\{\rho(\boldsymbol{w}_0^{\mathsf{T}}(\cdot) - b_0)\big\} = \delta(\boldsymbol{w}_0^{\mathsf{T}}(\cdot) - b_0)$$



- **Filtered Radon transform** of Laplacian of atom

$$\Longrightarrow \quad (\Lambda^{d-1}\,\mathscr{R}\,\Delta)\big\{\rho(\boldsymbol{w}_0^{\mathsf{T}}(\cdot) - b_0)\big\}(\boldsymbol{\gamma}, t) = \delta((\boldsymbol{\gamma}, t) - (\boldsymbol{w}_0, b_0)).$$

# Other activation functions

- Due to the **intertwining relations** of the Laplacian and the Radon transform, we can write

$$c_d \, \Lambda^{d-1} \, \mathscr{R} \, \Delta = c_d \, \partial_t^2 \, \Lambda^{d-1} \, \mathscr{R}$$

- The operator $\mathrm{R}_m := c_d \, \partial_t^m \Lambda^{d-1} \, \mathscr{R}$ sparsifies atoms of the form $\rho_m(\boldsymbol{w}_0^\mathsf{T}(\cdot) - b_0)$, where $\rho_m(x) = x_+^{m-1}/(m-1)!$.

  $\Longrightarrow$ In particular,

$$\mathrm{R}_m \, \rho_m(\boldsymbol{w}_0^\mathsf{T}(\cdot) - b_0) = \frac{\delta(\cdot - (\boldsymbol{w}_0, b_0)) + (-1)^m \delta(\cdot + (\boldsymbol{w}_0, b_0))}{2}$$

# Higher-order $\mathrm{BV}$ spaces in the Radon domain

- Recall

$$\mathrm{BV}^m(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} : \ \mathrm{D}^m f \in \mathcal{M}(\mathbb{R})\}$$

Consider the space

$$\mathscr{R}\,\mathrm{BV}^m(\mathbb{R}^d) := \left\{ f : \mathbb{R}^d \to \mathbb{R} : \begin{array}{l} c_d\, \partial_t^m \Lambda^{d-1}\, \mathscr{R}\, f \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R}), \\ \operatorname*{ess\,sup}_{\boldsymbol{x} \in \mathbb{R}^d} |f(\boldsymbol{x})| (1 + \|\boldsymbol{x}\|_2)^{-(m-1)} < \infty \end{array} \right\}$$

- When $d = 1$, one can show that $\mathscr{R}\,\mathrm{BV}^m(\mathbb{R}) = \mathrm{BV}^m(\mathbb{R})$.
  $\Longrightarrow$ Define $\mathscr{R}\,\mathrm{TV}^m(f) = c_d \|\partial_t^m \Lambda^{d-1}\, \mathscr{R}\, f\|_{\mathcal{M}}$.
  $\Longrightarrow$ When $d = 1$, $\mathscr{R}\,\mathrm{TV}^m(f) = \mathrm{TV}^m(f)$.

# Representer theorem for neural networks

## Theorem (P. and Nowak, 2021)

Consider the variational problem

$$\underset{f \in \mathscr{R}\,\mathrm{BV}^m(\mathbb{R}^d)}{\arg\min} \quad c_d \|\partial_t^m \Lambda^{d-1} \mathscr{R}\, f\|_{\mathcal{M}} \quad \text{s.t.} \quad \langle h_n, f \rangle = y_n,\ n = 1, \ldots, N,$$

where $h_n : \mathscr{R}\,\mathrm{BV}^m(\mathbb{R}^d) \to \mathbb{R}$ is weak$^*$ continuous, i.e., $h_n$ lies in the predual of $\mathscr{R}\,\mathrm{BV}^m(\mathbb{R}^d)$. Then, the solution set is nonempty, convex, and weak$^*$ compact. Moreover, the extreme points are shallow neural networks of the

$$s(\boldsymbol{x}) = \sum_{k=1}^{K} v_k\, \rho_m(\boldsymbol{w}_k^{\mathsf{T}} \boldsymbol{x} - b_k) + c(\boldsymbol{x}), \quad K \leq N,$$

where $c \in \mathcal{P}(\mathbb{R}^d)$ is a polynomial of degree $< m$.

# Representer theorem for neural networks

- Proof is similar to the proof for locally adaptive splines.
- Similar results hold for **deep neural networks**. (P. and Nowak 2021c)

  $\Longrightarrow$ Define $\mathscr{R}\,\mathrm{BV}^m(\mathbb{R}^d;\mathbb{R}^D)$ and consider compositions of such functions.

# What are $\mathscr{R}\,\mathrm{BV}^m$-spaces?

- Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with nice boundary (e.g., Lipschitz). Then,

$$\mathscr{R}\,\mathrm{BV}^m(\Omega) := \{f : \Omega \to \mathbb{R} : \ \exists g \in \mathscr{R}\,\mathrm{BV}^m(\mathbb{R}^d) \text{ s.t. } g\big|_\Omega = f\}$$

- When $d = 1$, $\mathscr{R}\,\mathrm{BV}^m(\Omega) = \mathrm{BV}^m(\Omega)$.
  $\implies$ $B^m_{1,1}(\Omega) \subset \mathrm{BV}^m(\Omega) \subset B^m_{1,\infty}(\Omega)$
  $\implies$ The **best** $K$-term approximation rate for $f \in \mathrm{BV}^m(\Omega)$ is

$$\|f - f_K\|_{L^2(\Omega)} \lesssim K^{-m},$$

  and is achieved by keeping the $K$ largest Daubechies wavelet coefficients or with free knot spline approximation.
- When $d > 1$,
  $\implies$ $W^{d-1+m,1}(\Omega) \subset \mathscr{R}\,\mathrm{BV}^m(\Omega)$. (P. and Nowak 2021b)
  $\implies$ We also know about the **best** approximation rates for $\mathscr{R}\,\mathrm{BV}^m(\Omega)$.

# Approximation properties of $\mathscr{R}\,\mathrm{BV}^m(\Omega)$

Suppose $\Omega = \left\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \leq 1\right\}$.

- Every $f \in \mathscr{R}\,\mathrm{BV}^m(\Omega)$ admits an integral representation

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} \rho_m(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)\,\mathrm{d}\mu(\boldsymbol{w}, b) + c(\boldsymbol{x}),$$

  where $\rho_m(x) = x_+^{m-1}/(m-1)!$, $\mu \in \mathcal{M}(\mathbb{S}^{d-1} \times [-1,1])$, and $c(\cdot)$ is a polynomial of degree $< m$.

# Dimension-free approximation rates

- Let $\mathcal{D} \coloneqq \{g\}_{g \in \mathcal{D}}$ be a dictionary of atoms such that $g \in L^2(\Omega)$. If $\mathcal{D} \subset L^2(\Omega)$ is compact, then there exists

$$f_K = \sum_{k=1}^{K} \alpha_k \, g_k, \quad g_k \in \mathcal{D}$$

such that

$$\left\| \int_{\mathcal{D}} g(\cdot)\mu(g(\cdot)) - f_K \right\|_{L^2(\Omega)} \lesssim K^{-\frac{1}{2}}$$

$\Longrightarrow$ This rate **does not grow** with the input dimension $d$.

$\Longrightarrow$ Gilles Pisier (1981). "Remarques sur un résultat non publié de B. Maurey". In: **Séminaire Analyse fonctionnelle (dit**, pp. 1–12

- In our problem, $\mathcal{D} = \left\{\rho_m(\boldsymbol{w}^\mathsf{T}(\cdot) - b)\right\}_{(\boldsymbol{w},b) \in \mathbb{S}^{d-1} \times [-1,1]}$.

# Approximation properties of $\mathscr{R}\,\mathrm{BV}^m(\Omega)$

- Given $f \in \mathscr{R}\,\mathrm{BV}^m(\Omega)$, there exists

$$f_K(\boldsymbol{x}) = \sum_{k=1}^{K} v_k\, \rho_m(\boldsymbol{w}_k^\mathsf{T} \boldsymbol{x} - b_k) + c(\boldsymbol{x})$$

such that

$$\|f - f_K\|_{L^2(\Omega)} \lesssim K^{-\frac{1}{2} - \frac{2m-1}{2d}} \lesssim K^{-\frac{1}{2}}.$$

This is the **best** rate. (Bach 2017; P. and Nowak 2021b; Siegel and Xu 2021)

- Compare this to the **best** $K$-term approximation rates in $H^m[0,1]^d$, which scales as

$$\|f - f_K\|_{L^2[0,1]^d} \lesssim K^{-\frac{m}{d}}$$

and is achieved by truncated Fourier series approximation.

$\Longrightarrow$ This rate **grows exponentially** with the input dimension $d$.

# Takeaway messages

- The intrinsic function spaces of neural networks are $\mathscr{R}\,\mathrm{BV}^m$-spaces.

  $\Longrightarrow$ These are **not** classically studied function spaces. Perhaps explaining the lack of understanding of neural networks in practice.

- These spaces are **small** in the sense of their "dimension-free" approximation rates compared to classical function spaces, e.g., Sobolev spaces.
- When $m = 1$, $\mathscr{R}\,\mathrm{TV}(\cdot) := \mathscr{R}\,\mathrm{TV}^1(\cdot)$ is a **new** notion of multivariate total variation, different than $\mathrm{TV}(f) = \|\nabla f\|_{\mathcal{M}}$.

  $\Longrightarrow$ When $m > 1$, $\mathscr{R}\,\mathrm{TV}^m(\cdot)$ provides a way to defined **higher-order** variants of multivariate total variation.

# Open problems

- How do $\mathscr{R}\,\mathrm{BV}^m$-spaces relate to classical spaces, e.g., Besov, Triebel-Lizorkin, etc.?

  $\Longrightarrow$ **Ridgelet analysis** is probably the right tool to answer this question.

- What other kinds of representer theorems can you get if you replace $\mathscr{R}$ with a **generalized Radon transform** in $c_d\,\partial_t^m\,\Lambda^{d-1}\,\mathscr{R}$.

  $\Longrightarrow$ Requires understanding the invertibility of the dual transform.

# References I



Bach, Francis (2017). "Breaking the curse of dimensionality with convex neural networks". In: **Journal of Machine Learning Research**.


Boyer, Claire, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric De Gournay, and Pierre Weiss (2019). "On representer theorems and convex regularization". In: **SIAM Journal on Optimization** 29.2, pp. 1260–1281.


Bredies, Kristian and Marcello Carioni (2020). "Sparsity of solutions for variational inverse problems with finite-dimensional data". In: **Calculus of Variations and Partial Differential Equations** 59.1, pp. 1–26.


Fisher, S. D. and Joseph W. Jerome (1975). "Spline solutions to $L^1$ extremal problems in one and several variables". In: **Journal of Approximation Theory**.


Ongie, Greg, Rebecca Willett, Daniel Soudry, and Nathan Srebro (2020). "A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case". In: **International Conference on Learning Representations**.


P., Rahul and Robert D. Nowak (2020). "The role of neural network activation functions". In: **IEEE Signal Processing Letters**.


— (2021a). "Banach Space Representer Theorems for Neural Networks and Ridge Splines". In: **Journal of Machine Learning Research**.


— (2021b). "Near-Minimax Optimal Estimation With Shallow ReLU Neural Networks". In: **arXiv preprint arXiv:2109.08844**.


— (2021c). "What Kinds of Functions do Deep Neural Networks Learn? Insights from Variational Spline Theory". In: **arXiv preprint arXiv:2105.03361**.

# References II



Pisier, Gilles (1981). "Remarques sur un résultat non publié de B. Maurey". In: **Séminaire Analyse fonctionnelle (dit**, pp. 1–12.


Siegel, Jonathan W. and Jinchao Xu (2021). "Characterization of the Variation Spaces Corresponding to Shallow Neural Networks". In: **arXiv preprint arXiv:2106.15002**.


Unser, Michael (2021). "A unifying representer theorem for inverse problems and machine learning". In: **Foundations of Computational Mathematics** 21.4, pp. 941–960.


Zuhovickiĭ, S. (1948). "Remarks on problems in approximation theory". In: **Mat. Zbirnik KDU**, pp. 169–183.