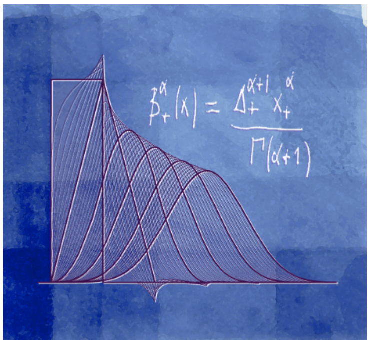
Part II: Variational Optimality of Neural Networks

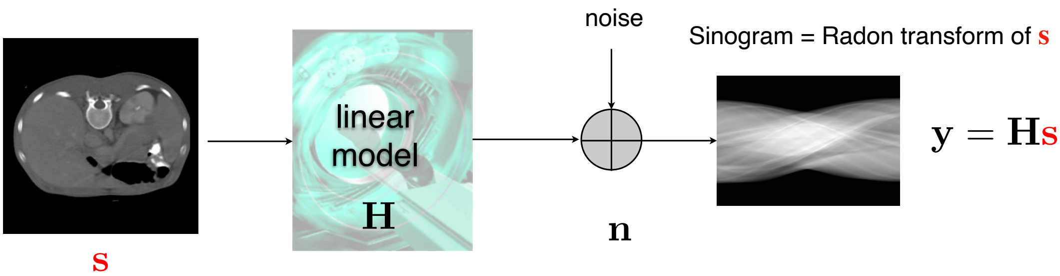
Michael Unser
Biomedical Imaging Group
EPFL, Lausanne, Switzerland



Summer School, Mathematics and Machine Learning for Image Analysis, Bologna, June 4-12, 2024

Variational formulation of inverse problems in imaging

- Linear forward model



$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}$$

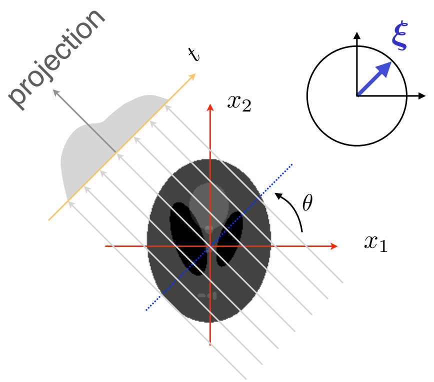
Problem: recover $\mathbf{s}$ from noisy measurements $\mathbf{y}$

- Regularization of ill-posed inverse problem

$$\mathbf{s}_{\text{rec}} = \arg\min_{\mathbf{s}\in\mathbb{R}^N} \underbrace{\|\mathbf{y} - \mathbf{H}\mathbf{s}\|_2^2}_{\text{data consistency}} + \underbrace{\lambda\|\mathbf{L}\mathbf{s}\|_p^p}_{\text{regularization}}, \quad p = 1, 2$$

The Radon transform and the **FBP algorithm**

Unit circle: $\mathbb{S}^1 = \{\boldsymbol{\xi} \in \mathbb{R}^2 : \|\boldsymbol{\xi}\| = 1\} = \{\boldsymbol{\xi} = (\cos\theta, \sin\theta), \theta \in [0, 2\pi)\}$



- Radon transform of $s \in L_1(\mathbb{R}^2)$

$$\mathrm{R}\{s\}(t, \boldsymbol{\xi}) = \int_{\boldsymbol{x} \in \mathbb{R}^2:\ \boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{x} = t} s(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$
$$= \int_{\mathbb{R}^2} \delta(t - \boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{x}) s(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \quad (t, \boldsymbol{\xi}) \in \mathbb{R} \times \mathbb{S}^1$$

- Reconstruction from $y(t, \boldsymbol{\xi}) = \mathrm{R}\{s\}(t, \boldsymbol{\xi})$: the **Filtered BackProjection** algorithm

$$s = \mathrm{R}^*\mathrm{K}_{\mathrm{rad}}\{y\}$$

  - $\mathrm{K}_{\mathrm{rad}}$: **"radial" filtering** in Radon space along the variable $t \in \mathbb{R}$.
    Fourier symbol $\widehat{K}_{\mathrm{rad}}(\omega) \propto |\omega|$
  - R^*: **backprojection** operator (the adjoint of R)

Supervised learning as a (linear) inverse problem

but an infinite-dimensional one …

Given the data points $(\boldsymbol{x}_m, y_m) \in \mathbb{R}^N \times \mathbb{R}$, find $f : \mathbb{R}^N \to \mathbb{R}$ s.t. $f(\boldsymbol{x}_m) \approx y_m$ for $m = 1, \ldots, M$

- Introduce smoothness or **regularization** constraint $(p = 2)$ (Poggio-Girosi 1990)

$R(f) = \|f\|_{\mathcal{H}}^2 = \|\mathrm{L}f\|_{L_2}^2 = \int_{\mathbb{R}^N} |\mathrm{L}f(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}$: regularization functional

$$\min_{f \in \mathcal{H}} R(f) \quad \text{subject to} \quad \sum_{m=1}^{M} |y_m - f(\boldsymbol{x}_m)|^2 \le \sigma^2$$

- Regularized least-squares fit (theory of RKHS)

$$f_{\mathrm{RKHS}} = \arg\min_{f \in \mathcal{H}} \left(\sum_{m=1}^{M} |y_m - f(\boldsymbol{x}_m)|^2 + \lambda R(f) \right) \quad \text{with} \quad R(f) = \|f\|_{\mathcal{H}}^2$$

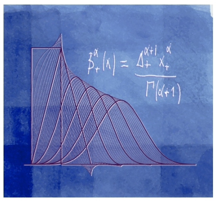
$\Rightarrow$ kernel estimator

(Wahba 1990; Schölkopf 2001)

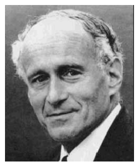
OUTLINE

- Connection with computational imaging ✔
- Variational formulation of learning: State-of-the art
  - Classical RKHS and kernel methods
  - Optimality results for shallow ReLU neural networks
- Radon-domain regularization for neural nets
  - Admissible regularization operator
  - Unifying representer theorem NEW
  - Laplacian revisited
  - Examples of admissible (operator, activation) pairs

(surprize?) ⇒ connection with *fractional* splines



Functions vs. distributions



Laurent Schwartz (1915-2002)

- Mathematical context
  - $\mathcal{S}(\mathbb{R}^d)$: Schwartz's space of smooth and rapidly-decaying functions on $\mathbb{R}^d$

    $\varphi : \mathbb{R}^d \to \mathbb{R}$ (or $\mathbb{C}$)

    $\boldsymbol{x} \mapsto \varphi(\boldsymbol{x})$
  - $\mathcal{S}'(\mathbb{R}^d)$: the space of **continuous linear functionals** on $\mathcal{S}(\mathbb{R}^d)$ = tempered distributions

    $f : \mathcal{S}(\mathbb{R}^d) \to \mathbb{R}$ (or $\mathbb{C}$)

    $\varphi \mapsto \langle f, \varphi \rangle = \int_{\mathbb{R}^d} f(\boldsymbol{x})\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ [*Formal* or *explicit* (for locally-integrable functions)]
- $L_2(\mathbb{R}^d)$: space of square-integrable functions on $\mathbb{R}^d$
  - L_2-norm: $\|f\|_{L_2} = \left(\int_{\mathbb{R}^d} |f(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}}$
  - $L_2(\mathbb{R}^d) = \{f : \mathbb{R}^d \to \mathbb{R}$ with f measurable and $\|f\|_{L_2} < \infty\}$ $= \overline{(\mathcal{S}(\mathbb{R}^d), \|\cdot\|_{L_2})}$
  - Continuous and dense embeddings: $\mathcal{S}(\mathbb{R}^d) \overset{\mathrm{d.}}{\hookrightarrow} L_2(\mathbb{R}^d) \overset{\mathrm{d.}}{\hookrightarrow} \mathcal{S}'(\mathbb{R}^d)$

RKHS representer theorem for L_2 regularization $(p = 2)$

(P2) $$\arg\min_{f \in \mathcal{H}} \left(\sum_{m=1}^{M} |y_m - f(\boldsymbol{x}_m)|^2 + \lambda \|f\|_{\mathcal{H}}^2 \right)$$ (deBoor 1966; Poggio-Girosi 1991)

$r_{\mathcal{H}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the (unique) **reproducing kernel** for the Hilbert $\mathcal{H} \subset \mathcal{S}'(\mathbb{R}^d)$ if
- $r_{\mathcal{H}}(\cdot, \boldsymbol{x}_0) \in \mathcal{H}$ for all $\boldsymbol{x}_0 \in \mathbb{R}^d$
- $f(\boldsymbol{x}_0) = \langle r_{\mathcal{H}}(\cdot, \boldsymbol{x}_0), f \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $\boldsymbol{x}_0 \in \mathbb{R}^d$

$\Leftrightarrow \quad \delta(\cdot - \boldsymbol{x}_0) \in \mathcal{H}'$

Convex loss function: $E : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$

Sample values: $\boldsymbol{f} = \big(f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_M)\big)$

(P2') $$\arg\min_{f \in \mathcal{H}} \left(E(\boldsymbol{y}, \boldsymbol{f}) + \lambda \|f\|_{\mathcal{H}}^2 \right)$$ (Schölkopf-Smola 2001)

Representer theorem for L_2-regularization

The generic parametric form of the solution of (P2') is

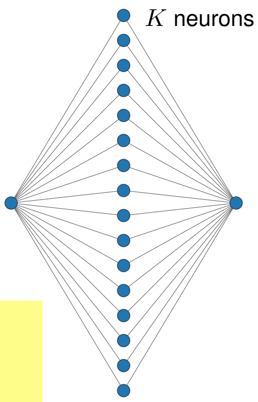
$$f(\boldsymbol{x}) = \sum_{m=1}^{M} a_m r_{\mathcal{H}}(\boldsymbol{x}, \boldsymbol{x}_m)$$

Supports the theory of SVM, kernel methods, variational splines, etc.

And what about neural networks ? Link with splines (gTV)

- **Shallow univariate** ReLU neural network with skip connection

$$f_{\boldsymbol{\theta}}(x) = c_0 + c_1 x + \sum_{k=1}^{K} v_k (w_k x - b_k)_+ \quad = c_0 + c_1 x + \sum_{k=1}^{K_0} a_k (x - \tau_k)_+$$



- Standard training with weight decay

(NN-1) : $$\arg \min_{\boldsymbol{\theta}=(\mathbf{v},\mathbf{w},\mathbf{b},\mathbf{c})} \sum_{m=1}^{M} \big|y_m - f_{\boldsymbol{\theta}}(x_m)\big|^2 + \frac{\lambda}{2} \sum_{k=1}^{K} |v_k|^2 + |w_k|^2$$

Theorem

For any $K \geq K_0$ (with $K_0 < M$), the solution of (NN-1) is achieved by the **sparse adaptive spline**:

$$f_{\text{spline}} = \arg \min_{f \in \mathrm{BV}^{(2)}(\mathbb{R})} \left(\sum_{m=1}^{M} |y_m - f(x_m)|^2 + \lambda \|\mathrm{D}^2 f\|_{\mathcal{M}} \right).$$

(U.-Fageot-Ward, 2017; Savarese 2019; Parhi-Nowak 2020)

Proper continuous counterpart of ℓ_1-norm

Johann Radon (1887-1956)

- Dual definition of ℓ_1-norm (in finite dimensions only)

$$\|\boldsymbol{f}\|_{\ell_1} = \sum_{n=1}^{N} |f_n| = \sup_{\boldsymbol{u}\in\mathbb{R}^N:\ \|\boldsymbol{u}\|_\infty \le 1} \langle \boldsymbol{f}, \boldsymbol{u}\rangle$$

- Space $C_0(\mathbb{R}^d)$ of functions on $\mathbb{R}^d$ that are continuous, bounded, and decaying at infinity

$$C_0(\mathbb{R}^d) = \overline{(\mathcal{S}(\mathbb{R}^d), \|\cdot\|_{L_\infty})} \ \subset \ L_\infty(\mathbb{R}^d)$$

- Space of **bounded Radon measures** on $\mathbb{R}^d$

$$\mathcal{M}(\mathbb{R}^d) = \big(C_0(\mathbb{R}^d)\big)' = \{f \in \mathcal{S}'(\mathbb{R}^d) : \|f\|_{\mathcal{M}} \triangleq \sup_{\varphi\in\mathcal{S}(\mathbb{R}^d):\ \|\varphi\|_\infty \le 1} \langle f, \varphi\rangle < +\infty\}$$

- **Superset** of $L_1(\mathbb{R}^d)$
  $\forall f \in L_1(\mathbb{R}^d): \quad \|f\|_{\mathcal{M}} = \|f\|_{L_1} \quad \Rightarrow \quad L_1(\mathbb{R}^d) \subset \mathcal{M}(\mathbb{R}^d)$

- **Extreme points** of unit ball in $\mathcal{M}(\mathbb{R}^d)$: $\quad e_k = \pm\delta(\cdot - \boldsymbol{\tau}_k)$ with $\boldsymbol{\tau}_k \in \mathbb{R}^d$

Multi-dimensional extension via hyper-spherical measures

- Integral representation of infinite-width shallow neural network

$$f(\boldsymbol{x}) = \int_{\mathbb{R}\times\mathbb{S}^{d-1}} \sigma(\boldsymbol{\xi}^\mathsf{T}\boldsymbol{x} - t)\mathrm{d}\mu(t,\boldsymbol{\xi})\mu(\cdot,\boldsymbol{\xi})\}(\boldsymbol{x}) \ = \mathrm{R}^*\{\sigma \circledast \mu(\cdot,\boldsymbol{\xi})\}(\boldsymbol{x})$$

R^*: Radon's backprojection operator

- Hyper-spherical counterpart of spike deconvolution problem (Duval-Peyré 2014; Bach 2017)

$$\arg \min_{\mu\in\mathcal{M}(\mathbb{R}\times\mathbb{S}^{d-1})} \left(\sum_{m=1}^{M} |y_m - \mathrm{R}^*\{\sigma \circledast \mu\}(\boldsymbol{x}_m)|^2 + \lambda\|\mu\|_{\mathcal{M}} \right)$$

Existence of solutions of the form: $f(\boldsymbol{x}) = \sum_{k=1}^{K_0} a_k \sigma(\boldsymbol{\xi}_k^\mathsf{T}\boldsymbol{x} - t_k)$

- Reproducing kernel Banach space (RKBS) (Bartolucci-DeVito-Rosasco-Vigogna; ACHA 2023)

$$\mathcal{B} = \{f_\mu : \mu \in \mathcal{M}(\Theta)\},$$
$$f_\mu(x) = \int_\Theta \rho(x,\theta)\beta(\theta)d\mu(\theta)$$
$$\|f\|_{\mathcal{B}} = \inf\{\|\mu\|_{\mathcal{M}} : f_\mu = f\}$$

$$\arg\min_{f\in\mathcal{B}} \left(\sum_{m=1}^{M} |y_m - f(x_m)|^2 + \lambda\|f\|_{\mathcal{B}} \right)$$

$$\Rightarrow f(x) = \sum_{k=1}^{K_0} a_k \rho(x, \theta_k)$$

OUTLINE

- Connection with computational imaging ✔
- Variational formulation of learning: State-of-the-art ✔
  - Classical RKHS and kernel methods
  - Optimality results for shallow ReLU neural networks
- Radon-domain regularization yields neural nets
  - Admissible regularization operators
  - Null space of polynomials
  - Unifying representer theorem
  - Native spaces
  - Example of admissible (operator, activation) pairs

Admissible regularization operator

- Isotropic convolution operator

A linear operator $\mathrm{L} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ that is **shift-invariant** and **isotropic** is uniquely characterized by its **radial frequency profile** $\widehat{L}_{\mathrm{rad}} : \mathbb{R} \to \mathbb{R}$.

Fourier symbol of L: $\widehat{L}(\boldsymbol{\omega}) = \widehat{L}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|)$

Definition
An isotropic regularization operator with frequency profile $\widehat{L}_{\mathrm{rad}}(\omega)$ is **spline-admissible** with a **polynomial null space** of degree n_0 (possibly trivial) if

1. $\widehat{L}_{\mathrm{rad}}(\omega)$ **does not vanish** over $\mathbb{R}$, except for **a zero** of order $\gamma_0 \in (n_0, n_0+1]$ **at the origin**; that is, $|\widehat{L}_{\mathrm{rad}}(\omega)|/|\omega|^{\gamma_0} = C_0$ as $\omega \to 0$.
2. Ellipticity: There exists an order $\gamma_1 > 1$, a constant $C_1 > 0$, and a radius $R_1 > 0$ such that $|\widehat{L}_{\mathrm{rad}}(\omega)| \geq C_1|\omega|^{\gamma_1}$ for all $|\omega| > R_1$.

- Example

$\mathrm{L} = \Delta$ (Laplacian) with $\widehat{L}_{\mathrm{rad}}(\omega) = -\omega^2$, $\gamma_0 = \gamma_1 = 2$, and $n_0 = 1$.

Null space of polynomials

Isotropic regularization operator L with frequency profile $\widehat{L}_{\mathrm{rad}} : \mathbb{R} \to \mathbb{R}$

- Effect of a γ_0th-order zero

  $\lim_{\omega \to 0} \frac{|\widehat{L}_{\mathrm{rad}}(\omega)|}{|\omega|^{\gamma_0}} = C_0 \quad \Rightarrow$ annihilates polynomials of degree $n_0 = \lceil \gamma_0 - 1 \rceil$

- The space $\mathcal{P}_{n_0}$ of polynomials of degree n_0
  - Taylor (or monomial) basis: $m_{\boldsymbol{k}}(\boldsymbol{x}) \triangleq \frac{\boldsymbol{x}^{\boldsymbol{k}}}{\boldsymbol{k}!}$
  - $\mathcal{P}_{n_0} = \{p_0 = \sum_{|\boldsymbol{k}| \le n_0} b_{\boldsymbol{k}} m_{\boldsymbol{k}} : \|p_0\|_{\mathcal{P}} \triangleq \|(b_{\boldsymbol{k}})_{|\boldsymbol{k}| \le n_0}\|_2 < \infty\}$.

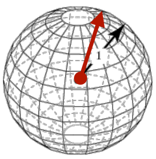
Proposition (Construction of biothogonal basis)
There exists an isotropic window $\kappa_{\mathrm{iso}} \in \mathcal{S}(\mathbb{R}^d)$ with $0 \le \widehat{\kappa}_{\mathrm{iso}}(\boldsymbol{\omega}) \le 1$ and $\widehat{\kappa}_{\mathrm{iso}}(\boldsymbol{\omega}) = 0$ for $\|\boldsymbol{\omega}\| \ge 1$ such that, for all $\boldsymbol{k}, \boldsymbol{n} \in \mathbb{N}^d$,

$$m^*_{\boldsymbol{n}} \triangleq (-1)^{|\boldsymbol{n}|} \partial^{\boldsymbol{n}} \kappa_{\mathrm{iso}} \quad \text{and} \quad \langle m_{\boldsymbol{k}}, m^*_{\boldsymbol{n}} \rangle = \delta_{\boldsymbol{k}-\boldsymbol{n}} \quad \text{(biorthogonality)}$$

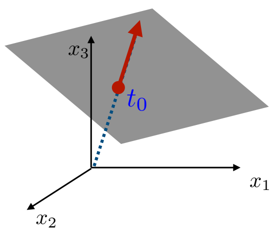
- Dual space $\mathcal{P}'_{n_0} = \left\{p^*_0 = \sum_{|\boldsymbol{k}| \le n_0} b^*_{\boldsymbol{k}} m^*_{\boldsymbol{k}} : \|p^*_0\|_{\mathcal{P}'} \triangleq \|(b^*_{\boldsymbol{k}})\|_2 < \infty\right\} \subset \mathcal{S}(\mathbb{R}^d)$

The Radon transform: Classical integral formulation

Unit sphere: $\mathbb{S}^{d-1} = \{\boldsymbol{\xi} \in \mathbb{R}^d : \|\boldsymbol{\xi}\| = 1\}$



Hyperplane $P_{\boldsymbol{\xi}_0, t_0} = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{\xi}_0{}^{\mathsf{T}} \boldsymbol{x} = t_0\}$



- Radon transform of $f \in L_1(\mathbb{R}^d)$

$$\begin{aligned} \mathrm{R}\{f\}(t, \boldsymbol{\xi}) &= \int_{\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{x} = t} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= \int_{\mathbb{R}^d} \delta(t - \boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \quad (t, \boldsymbol{\xi}) \in \mathbb{R} \times \mathbb{S}^{d-1} \end{aligned}$$

- Backprojection operator: From Radon domain to Euclidean space

$$\mathrm{R}^*\{g\}(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}} g(\underbrace{\boldsymbol{\xi}^{\mathsf{T}} \boldsymbol{x}}_{t}, \boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}, \quad \boldsymbol{x} \in \mathbb{R}^d$$

- Fourier slice theorem

$$\mathrm{R}\{f\}(t, \boldsymbol{\xi}_0) = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\omega \boldsymbol{\xi}_0) \mathrm{e}^{\mathrm{i}\omega t} \mathrm{d}\omega = \mathcal{F}^{-1}_{\omega \to t}\{\hat{f}(\omega \boldsymbol{\xi}_0)\}\{t\}$$

Fourier transform:

$$\hat{f}(\boldsymbol{\omega}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} f(\boldsymbol{x}) \mathrm{e}^{-\mathrm{i}\langle \boldsymbol{\omega}, \boldsymbol{x} \rangle} \mathrm{d}\boldsymbol{x}$$

New twist: Generic, Radon-domain regularization

Isotropic regularization operator L with frequency profile $\widehat{L}_{\mathrm{rad}} : \mathbb{R} \to \mathbb{R}$

"Radonized" version of the operator:

$$\mathrm{L}_{\mathrm{R}} = \mathrm{K}_{\mathrm{rad}}\mathrm{R}\mathrm{L}$$

- Role of each operator
  - L: differential operator such as Laplacian, to penalize high frequency components
  - R: Radon transform, to project in hyperspherical domain with $(t, \boldsymbol{\xi}) \in \mathbb{R} \times \mathbb{S}^{d-1}$
  - $\mathrm{K}_{\mathrm{rad}}$: isotropic Radon-domain filtering with $\widehat{K}_{\mathrm{rad}}(\omega) \propto |\omega|^{d-1}$, to facilitate inversion

- "Easy" case where L is invertible = trivial null space
  - $\mathrm{L}^{-1}\mathrm{L} = \mathrm{Id}$ on $\mathcal{S}'(\mathbb{R}^d)$
  - L_{R} has a trivial null space
  - Inversion of Radon transform: $\mathrm{R}^*\mathrm{K}_{\mathrm{rad}}\mathrm{R} = \mathrm{Id}$ on $\mathcal{S}'(\mathbb{R}^d)$

  $\Rightarrow \quad \mathrm{L}_{\mathrm{R}}^{-1} = \mathrm{L}^{-1}\mathrm{R}^*$ (Ludwig 1966)

- Non-trivial null space
  - L and L_{R} share the same null space: $\mathcal{P}_{n_0}$
  - Canonical scenario: $n_0 = 1$ (affine maps)
  - Makes the inversion process more difficult

Representer theorem for neural nets: Context

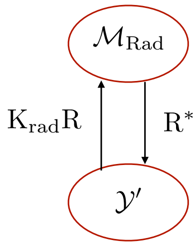
Given the data points $(\boldsymbol{x}_m, y_m) \in \mathbb{R}^d \times \mathbb{R}$, find $f : \mathbb{R}^d \to \mathbb{R}$ s.t. $f(\boldsymbol{x}_m) \approx y_m$ for $m = 1, \dots, M$

- Variational formulation with Radon-domain regularization

$$S = \arg \min_{f \in \mathcal{M}_{\mathrm{L}_{\mathrm{R}}}(\mathbb{R}^d)} \sum_{m=1}^{M} E\big(y_m, f(\boldsymbol{x}_m)\big) + \psi\left(\|\mathrm{L}_{\mathrm{R}} f\|_{\mathcal{M}(\mathbb{R}\times\mathbb{S}^{d-1})}\right)$$

  - **Regularization operator**: $\mathrm{L}_{\mathrm{R}} = \mathrm{K}_{\mathrm{rad}}\mathrm{R}\mathrm{L} : \mathcal{M}_{\mathrm{L}_{\mathrm{R}}}(\mathbb{R}^d) \to \mathcal{M}_{\mathrm{Rad}}$ where L is admissible
  - **Native space**: $\mathcal{M}_{\mathrm{L}_{\mathrm{R}}}(\mathbb{R}^d)$ = Banach space that is isometrically isomorphic to $\mathcal{M}_{\mathrm{Rad}} \times \mathcal{P}_{n_0}$
  - $\mathcal{M}_{\mathrm{Rad}} = \mathcal{M}_{\mathrm{even}}(\mathbb{R} \times \mathbb{S}^{d-1})$: Banach space of **Radon-compatible bounded measures**
  - $E : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is a strictly-convex loss functional.
  - $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ is some arbitrary strictly-increasing convex function.

Theoretical ingredients



- Specification of Euclidean-to-Radon-domain isomorphisms
  - Problem: the Radon transform is not surjective on $\mathcal{S}$ (resp., $\mathcal{S}'$)
  - Identification of proper hyperspherical Banach subspace $\mathcal{M}_{\mathrm{Rad}} \subset \mathcal{M}(\mathbb{R} \times \mathbb{S}^{d-1})$ over which R^* is invertible

$\Rightarrow$ **Banach-space variant** of the distributional theory of the Radon transform

(Unser, JMLR 2023)

- Dealing with the non-trivial null space of L_{R}
  - Take inspiration from spline theory (U.-Fageot-Ward, SIAM Rev 2017)

  Factoring out the null space: $(\mathrm{Id} - \mathrm{Proj}_{\mathcal{P}_{n_0}})\{f\} = f - \sum_{|\boldsymbol{k}| \le n_0} \langle f, m_{\boldsymbol{k}}^* \rangle \, m_{\boldsymbol{k}}$

- Abstract representer theorem for Banach space with semi-norm penalties (U.-Aziznejad, ACHA 2022)
  - Prove that $\mathcal{M}_{\mathrm{L}_{\mathrm{R}}}$ is a Banach space (Unser FoCM in press)
  - Establish weak* continuity of sampling functionals

Representer theorem for neural nets

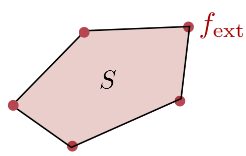
$$S = \arg \min_{f \in \mathcal{M}_{\mathrm{L}_{\mathrm{R}}}(\mathbb{R}^d)} \sum_{m=1}^{M} E\big(y_m, f(\boldsymbol{x}_m)\big) + \psi\left(\|\mathrm{L}_{\mathrm{R}} f\|_{\mathcal{M}(\mathbb{R}\times\mathbb{S}^{d-1})}\right) \qquad (1)$$

$$\mathrm{L}_{\mathrm{R}} = \mathrm{K}_{\mathrm{rad}}\mathrm{R}\mathrm{L} : \mathcal{M}_{\mathrm{L}_{\mathrm{R}}}(\mathbb{R}^d) \to \mathcal{M}(\mathbb{R} \times \mathbb{S}^{d-1})$$

Theorem
The solution set S of Problem (1) is non-empty and weak* compact. It is the weak* closure of the convex hull of its **extreme points**, which can all be written as

$$f_{\mathrm{ext}}(\boldsymbol{x}) = p_0(\boldsymbol{x}) + \sum_{k=1}^{K_0} a_k \rho_{\mathrm{rad}}(\boldsymbol{\xi}_k^{\mathsf{T}}\boldsymbol{x} - \tau_k)$$

with a fixed **activation function** $\rho_{\mathrm{rad}} = \mathcal{F}^{-1}\{1/\widehat{L}_{\mathrm{rad}}\}$, for some $K_0 \le M - \dim \mathcal{P}_{n_0}$, $(a_k, \boldsymbol{\xi}_k, \tau_k) \in \mathbb{R} \times \mathbb{S}^{d-1} \times \mathbb{R}$ for $k = 1, \dots, K_0$, and a **null-space component** $p_0 \in \mathcal{P}_{n_0}$. The corresponding regularization cost (shared by all solutions) is $\|\mathrm{L}_{\mathrm{R}} f_{\mathrm{ext}}\|_{\mathcal{M}(\mathbb{R}\times\mathbb{S}^{d-1})} = \sum_{n=1}^{K_0} |a_k|$.



Special case of abstract rep theorem for direct sums (U.-Aziznejad, ACHA 2022)

Special case: Laplacian

- Properties of the (negative) Laplacian
  - $(-\Delta)f(\boldsymbol{x}) = -\sum_{n=1}^{d} \frac{\partial^2}{\partial x_n^2} f(\boldsymbol{x})$
  - Frequency symbol: $\|\boldsymbol{\omega}\|^2 \quad \Rightarrow \quad \widehat{\Delta}_{\mathrm{rad}}(\omega) = \omega^2$ (radial profile)
  - Annihilates all affine functions: $n_0 = 1$

- Outcome of representer theorem with $\mathrm{L_R} = \mathrm{K_{rad}R}(-\Delta)$
  - Null space: $\mathcal{P}_1 = \{p_0(\boldsymbol{x}) = b_0 + \boldsymbol{b}^{\mathsf{T}}\boldsymbol{x} : (b_0, \boldsymbol{b}) \in \mathbb{R}^{d+1}\}$
  - Activation function: $\rho_{\mathrm{rad}}(t) = \mathcal{F}^{-1}\{\frac{1}{\omega^2}\}(t) = \frac{1}{2}|t| = t_+ - \frac{1}{2}t$

$\Rightarrow$ Shallow ReLU net: $f_{\mathrm{ext}}(\boldsymbol{x}) = c_0 + \boldsymbol{c}^{\mathsf{T}}\boldsymbol{x} + \sum_{k=1}^{K_0} a_k(\boldsymbol{\xi}_k^{\mathsf{T}}\boldsymbol{x} - \tau_k)_+$

sum of elementary (ReLU) ridges = **ridge spline** (Parhi-Nowak 2021)

Limit behaviour of multivariate 2-layer ReLU neural nets

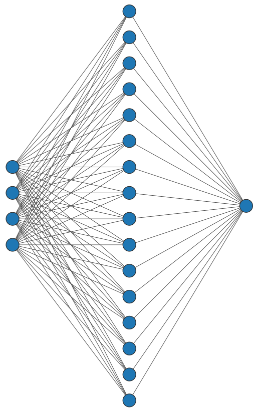
- Shallow ReLU neural network $\mathbb{R}^d \to \mathbb{R}$ with skip connection

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = c_0 + \boldsymbol{c}_1^{\mathsf{T}}\boldsymbol{x} + \sum_{k=1}^{K} v_k(\boldsymbol{w}_k^{\mathsf{T}}\boldsymbol{x} - b_k)_+ \qquad = c_0 + \boldsymbol{c}_1^{\mathsf{T}}\boldsymbol{x} + \sum_{k=1}^{K_0} a_k(\boldsymbol{\xi}_k^{\mathsf{T}}\boldsymbol{x} - \tau_k)_+$$

- Standard training with weight decay on $\mathbf{v} = (v_k)$ and $\mathbf{W} = [\boldsymbol{w}_1 \ \dots \boldsymbol{w}_K]$

$$(\text{NN-}d): \quad \arg \min_{\boldsymbol{\theta}=(\mathbf{v},\mathbf{W},\mathbf{b},\mathbf{c})} \sum_{m=1}^{M} \left|y_m - f_{\boldsymbol{\theta}}(\boldsymbol{x}_m)\right|^2 + \frac{\lambda}{2}\sum_{k=1}^{K} |v_k|^2 + \|\boldsymbol{w}_k\|^2$$

K neurons



Theorem
For any $K \geq K_0$ (with $K_0 < M$), the solution of (NN-d) is achieved by the **sparse ridge spline**:

$$f_{\mathrm{ridge}} = \arg \min_{f \in \mathcal{M}_{\Delta_{\mathrm{R}}}(\mathbb{R}^d)} \left(\sum_{m=1}^{M} |y_m - f(\boldsymbol{x}_m)|^2 + \lambda \|\mathrm{K_{rad}R}\Delta f\|_{\mathcal{M}(\mathbb{R}\times\mathbb{S}^{d-1})} \right).$$

(Ongie et al. 2020; Parhi-Nowak 2021)

Delicate point: Proper delineation of the native space $\mathcal{M}_{\Delta_{\mathrm{R}}}(\mathbb{R}^d)$

NATIVE SPACES

Hyper-spherical (test) functions and distributions

- Test functions and tempered distributions

$$(\phi, g) \in \mathcal{S}(\mathbb{R} \times \mathbb{S}^{d-1}) \times \mathcal{S}'(\mathbb{R} \times \mathbb{S}^{d-1})$$

$$g : \phi \mapsto \langle g, \phi \rangle_{\mathrm{Rad}} \in \mathbb{R}$$

For locally integrable functions $g : (t, \boldsymbol{\xi}) \mapsto \mathbb{R}$:

$$\langle g, \phi \rangle_{\mathrm{Rad}} = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} g(t, \boldsymbol{\xi}) \phi(t, \boldsymbol{\xi}) \mathrm{d}t \mathrm{d}\boldsymbol{\xi}$$

Special case: $d = 2$

$\boldsymbol{\xi} = (\cos\theta, \sin\theta)$ with $\mathrm{d}\boldsymbol{\xi} = \mathrm{d}\theta$ for $\theta \in [0, 2\pi]$

$$\langle g, \phi \rangle_{\mathrm{Rad}} = \int_0^{2\pi} \int_{\mathbb{R}} g(t, \theta) \phi(t, \theta) \mathrm{d}t \mathrm{d}\theta$$

- Radon transform and its adjoint

$$\mathrm{R} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R} \times \mathbb{S}^{d-1})$$

$$\mathrm{R}^* : \mathcal{S}'(\mathbb{R} \times \mathbb{S}^{d-1}) \to \mathcal{S}'(\mathbb{R}^d)$$

R^* is the unique linear operator such that

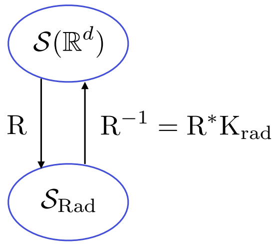
$$\forall \varphi \in \mathcal{S}(\mathbb{R}^d) : \quad \langle \mathrm{R}^*\{g\}, \varphi \rangle = \langle g, \mathrm{R}\{\varphi\} \rangle_{\mathrm{Rad}}$$

Radon transform on $\mathcal{S}(\mathbb{R}^d)$

Theorem (Invertibility of Radon transform on $\mathcal{S}(\mathbb{R}^d)$)

1. R continuously maps $\mathcal{S}(\mathbb{R}^d) \to \mathcal{S}_{\mathrm{Rad}} \subset \mathcal{S}(\mathbb{R} \times \mathbb{S}^{d-1})$
2. $\mathrm{R}^*\mathrm{K}_{\mathrm{rad}}\mathrm{R} = \mathrm{Id}$ on $\mathcal{S}(\mathbb{R}^d)$ $\quad \Leftrightarrow \quad$ $\mathrm{R}^{-1} = \mathrm{R}^*\mathrm{K}_{\mathrm{rad}}$ on $\mathcal{S}_{\mathrm{Rad}} = \mathrm{R}\big(\mathcal{S}(\mathbb{R}^d)\big)$

(Gelfand 1962; Helgason 1965; Ludwig 1966)



- Radon-domain filtering operator
  - $\mathrm{K}_{\mathrm{rad}}$: "radial" operator that acts along the Radon-domain variable t
  - Radial frequency response: $\widehat{K}_{\mathrm{rad}}(\omega) = c_d|\omega|^{d-1}$

Distributional theory of the (filtered) Radon transform

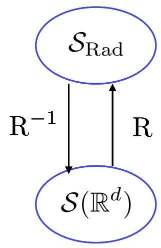
$\mathrm{R}: \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R} \times \mathbb{S}^{d-1})$

Difficulty: injective but not surjective !

Theorem (Variant on invertibility of Radon transform)

$\mathcal{S}_{\mathrm{Rad}} = \mathrm{R}\big(\mathcal{S}(\mathbb{R}^d)\big)$ is a closed subspace of $\mathcal{S}(\mathbb{R} \times \mathbb{S}^{d-1})$. Moreover,

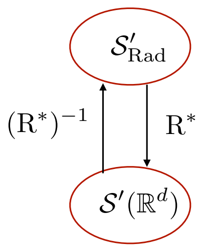
1. $\mathrm{R}: \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}_{\mathrm{Rad}}$ is a **continuous bijection**, with $\mathrm{R}^*\mathrm{K}_{\mathrm{rad}}\mathrm{R} = \mathrm{Id}$ on $\mathcal{S}(\mathbb{R}^d)$
2. $\mathrm{R}^*\mathrm{K}_{\mathrm{rad}}: \mathcal{S}_{\mathrm{Rad}} \to \mathcal{S}(\mathbb{R}^d)$ is a **continuous bijection**, with $\mathrm{R}\mathrm{R}^*\mathrm{K}_{\mathrm{rad}} = \mathrm{Id}$ on $\mathcal{S}_{\mathrm{Rad}}$
3. $\mathrm{R}^*: \mathcal{S}'_{\mathrm{Rad}} \to \mathcal{S}'(\mathbb{R}^d)$ is a **continuous bijection** with $\mathrm{K}_{\mathrm{rad}}\mathrm{R}\mathrm{R}^* = \mathrm{Id}$ on $\mathcal{S}'_{\mathrm{Rad}}$.



Theorem (Characterization of the range space)
Let $\phi \in \mathcal{S}(\mathbb{R} \times \mathbb{S}^{d-1})$. Then, $\phi \in \mathcal{S}_{\mathrm{Rad}} \triangleq \big\{\phi = \mathrm{R}\{\varphi\} : \varphi \in \mathcal{S}(\mathbb{R}^d)\big\}$ iff.

1. Evenness: $\phi(t, \boldsymbol{\xi}) = \phi(-t, -\boldsymbol{\xi})$.
2. $\Phi_k(\boldsymbol{\xi}) = \int_{\mathbb{R}} \phi(t, \boldsymbol{\xi}) t^k \mathrm{d}t$ is a homogeneous polynomial in $\boldsymbol{\xi} \in \mathbb{S}^{d-1}$ for any $k \in \mathbb{N}_0$

(Gelfand 1962; Helgason 1965; Ludwig 1966)



Banach space theory of the (filtered) Radon transform

$$\mathcal{X} = \overline{\left(\mathcal{S}(\mathbb{R}\times\mathbb{S}^{d-1}), \|\cdot\|_{\mathcal{X}}\right)} \ \subset\ \mathcal{S}'(\mathbb{R}\times\mathbb{S}^{d-1})$$

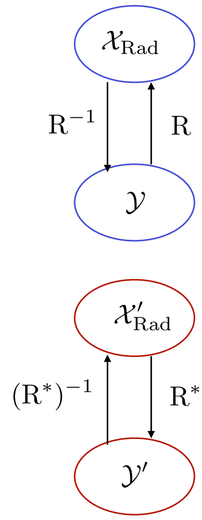
$$\mathcal{X}' = \left\{g \in \mathcal{S}'(\mathbb{R}\times\mathbb{S}^{d-1}) : \|g\|_{\mathcal{X}'} < \infty\right\} \quad \text{with} \quad \|g\|_{\mathcal{X}'} = \sup_{\phi\in\mathcal{S}(\mathbb{R}\times\mathbb{S}^{d-1}):\ \|\phi\|_{\mathcal{X}}\le 1} \langle g, \phi\rangle_{\mathrm{Rad}}$$

$$\mathcal{X}_{\mathrm{Rad}} = \overline{(\mathcal{S}_{\mathrm{Rad}}, \|\cdot\|_{\mathcal{X}})} \subset \mathcal{X}$$

Theorem (Radon-compatible Banach isometries)

Let $\|\varphi\|_{\mathcal{Y}} \triangleq \|\mathrm{R}\{\varphi\}\|_{\mathcal{X}}$. Then, $\mathrm{R} : (\mathcal{S}(\mathbb{R}^d), \|\cdot\|_{\mathcal{Y}}) \to (\mathcal{S}_{\mathrm{Rad}}, \|\cdot\|_{\mathcal{X}})$ has a **unique isometric extension** $\mathrm{R} : \mathcal{Y} \to \mathcal{X}_{\mathrm{Rad}}$ with $\mathcal{Y} = \overline{(\mathcal{S}(\mathbb{R}^d), \|\cdot\|_{\mathcal{Y}})}$. Moreover,

1. $\mathrm{R}^*\mathrm{K}_{\mathrm{rad}} : \mathcal{X}_{\mathrm{Rad}} \to \mathcal{Y}$ is an **isometric bijection**, with $\mathrm{R}\mathrm{R}^*\mathrm{K}_{\mathrm{rad}} = \mathrm{Id}$ on $\mathcal{X}_{\mathrm{Rad}}$
2. $\mathrm{R}^* : \mathcal{X}'_{\mathrm{Rad}} \to \mathcal{Y}'$ is an **isometric bijection** with $\mathrm{K}_{\mathrm{rad}}\mathrm{R}\mathrm{R}^* = \mathrm{Id}$ on $\mathcal{X}'_{\mathrm{Rad}}$.



(Unser, JMLR 2022)

Hyper-spherical functions and measures

- Banach space of hyper-spherical bounded Radon measures

$$\mathcal{M}(\mathbb{R}\times\mathbb{S}^{d-1}) = \left(C_0(\mathbb{R}\times\mathbb{S}^{d-1})\right)' \quad \text{where} \quad C_0(\mathbb{R}\times\mathbb{S}^{d-1}) = \overline{\left(\mathcal{S}(\mathbb{R}\times\mathbb{S}^{d-1}), \|\cdot\|_{L_\infty}\right)}$$

Null space of R^*: $\ker(\mathrm{R}^*) = \left\{g \in \mathcal{M}(\mathbb{R}\times\mathbb{S}^{d-1}) : \langle \mathrm{R}\{\varphi\}, g\rangle_{\mathrm{Rad}} = \langle \varphi, \mathrm{R}^*\{g\}\rangle = 0, \forall \varphi \in \mathcal{S}(\mathbb{R}^d)\right\}$

Theorem (Inversion of backprojection operator) (Neumayer-U., Anal. and Appl. 2023)

The **quotient space** $\mathcal{M}_{\mathrm{Rad}} = \mathcal{M}(\mathbb{R}\times\mathbb{S}^{d-1})/\ker(\mathrm{R}^*)$ is a Banach space that is **isometrically isomorphic** to $\mathcal{M}_{\mathrm{even}}(\mathbb{R}\times\mathbb{S}^{d-1})$. Consequently, $\mathrm{K}_{\mathrm{rad}}\mathrm{R}\mathrm{R}^* = \mathrm{Id}$ on $\mathcal{M}_{\mathrm{even}}$ (resp., $\mathcal{M}_{\mathrm{Rad}}$).

$$C_{0,\mathrm{even}}(\mathbb{R}\times\mathbb{S}^{d-1}) = \overline{(\mathcal{S}_{\mathrm{Rad}}, \|\cdot\|_{L_\infty})} \subset C_0(\mathbb{R}\times\mathbb{S}^{d-1})$$

$$L_{p,\mathrm{even}}(\mathbb{R}\times\mathbb{S}^{d-1}) = \overline{(\mathcal{S}_{\mathrm{Rad}}, \|\cdot\|_{L_p})} \subset L_p(\mathbb{R}\times\mathbb{S}^{d-1}), \quad p \in (1,\infty)$$

Native space for Radon-domain regularization

- Regularization functional: $\|\mathrm{L_R}f\|_{\mathcal{M}}$ with $\mathrm{L_R} = \mathrm{K_{rad}RL} : \mathcal{M}_{\mathrm{L_R}}(\mathbb{R}^d) \to \mathcal{M}_{\mathrm{Rad}}$
  - Radon-domain $\mathcal{M}$-norm : $\|g\|_{\mathcal{M}} \triangleq \sup_{\phi \in \mathcal{S}(\mathbb{R}\times\mathbb{S}^{d-1}):\|\phi\|_{L_\infty}\leq 1} \langle g, \phi\rangle_{\mathrm{Rad}},$

- Native space

$$\begin{aligned}\mathcal{M}_{\mathrm{L_R}}(\mathbb{R}^d) &= \mathrm{L_R^\dagger}\big(\mathcal{M}_{\mathrm{Rad}}\big) \oplus \mathcal{P}_{n_0}\\ &= \{f = \mathrm{L_R^\dagger}\{w\} + p_0 : (w, p_0) \in \mathcal{M}_{\mathrm{Rad}} \times \mathcal{P}_{n_0}\},\end{aligned}$$

  - **Null space** of L and $\mathrm{L_R}$: $\mathcal{P} = \mathcal{P}_{n_0} = \mathrm{span}\{m_{\boldsymbol{k}}\}_{|\boldsymbol{k}|\leq n_0}$ with $m_{\boldsymbol{k}}(\boldsymbol{x}) = \frac{\boldsymbol{x}^{\boldsymbol{k}}}{\boldsymbol{k}!}$
  - **Right-inverse** of $\mathrm{L_R}$ on $\mathcal{M}_{\mathrm{Rad}}$: $\mathrm{L_R^\dagger} = (\mathrm{Id} - \mathrm{Proj}_{\mathcal{P}})\mathrm{L}^{-1}\mathrm{R}^*$
  - **Projector** $\mathrm{Proj}_{\mathcal{P}} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{P} : f \mapsto \sum_{|\boldsymbol{k}|\leq n_0} \langle f, m^*_{\boldsymbol{k}}\rangle\, m_{\boldsymbol{k}}$ with $m^*_{\boldsymbol{k}} = (-1)^{|\boldsymbol{k}|}\partial^{\boldsymbol{k}}\kappa_{\mathrm{iso}} \in \mathcal{S}(\mathbb{R}^d)$

Properties of native space

Right-inverse property: $\mathcal{U}' = \mathrm{L_R^\dagger}(\mathcal{M}_{\mathrm{Rad}}) \quad \Leftrightarrow \quad \mathrm{L_R}(\mathcal{U}') = \mathcal{M}_{\mathrm{Rad}}$

Theorem (Unser FoCM in press)

Let L be an admissible operator with a polynomial null space $\mathcal{P} = \mathcal{P}_{n_0}$ (possibly trivial) of degree n_0.

1. The space $\mathcal{M}_{\mathrm{L_R}}(\mathbb{R}^d) = \mathcal{U}' \oplus \mathcal{P}$ equipped with the composite norm $\|f\|_{\mathcal{M}_{\mathrm{L_R}}} = \|\mathrm{L_R}\{f\}\|_{\mathcal{M}} + \|\mathrm{Proj}_{\mathcal{P}}\{f\}\|_{\mathcal{P}}$ is **complete** and **isomorphic** to $\mathcal{M}_{\mathrm{Rad}} \times \mathcal{P}$

2. The operators $\mathrm{L_R} = \mathrm{K_{rad}RL} : \mathcal{M}_{\mathrm{L_R}} \to \mathcal{M}_{\mathrm{Rad}}$ and $\mathrm{L_R^\dagger} = (\mathrm{Id} - \mathrm{Proj}_{\mathcal{P}})\mathrm{L}^{-1}\mathrm{R}^* : \mathcal{M}_{\mathrm{Rad}} \to L_{\infty,-n_0}(\mathbb{R}^d)$ are continuous and have the following properties:

$$\begin{aligned}\forall w \in \mathcal{M}_{\mathrm{Rad}} :&\quad \mathrm{L_R}\mathrm{L_R^\dagger}\{w\} = w\\ \forall p_0 \in \mathcal{P} :&\quad \mathrm{L_R}\{p_0\} = 0\\ \forall f \in \mathcal{M}_{\mathrm{L_R}}(\mathbb{R}^d) :&\quad \mathrm{L_R^\dagger}\mathrm{L_R}\{f\} = (\mathrm{Id} - \mathrm{Proj}_{\mathcal{P}})\{f\} = \mathrm{Proj}_{\mathcal{U}'}\{f\}.\end{aligned}$$

3. Embeddings: $\mathcal{S}(\mathbb{R}^d) \hookrightarrow \mathcal{M}_{\mathrm{L_R}}(\mathbb{R}^d) \hookrightarrow L_{\infty,-n_0}(\mathbb{R}^d) \overset{\mathrm{d.}}{\hookrightarrow} \mathcal{S}'(\mathbb{R}^d)$.

$L_{\infty,-n_0}(\mathbb{R}^d) = \{f : \mathbb{R}^d \to \mathbb{R} \text{ s.t. } \sup_{\boldsymbol{x}\in\mathbb{R}^d}(1 + \|\boldsymbol{x}\|)^{-n_0}|f(\boldsymbol{x})| < \infty\}$

Schwartz kernel of pseudo-inverse operator

- Adjoint pair of pseudo-inverse operators

$$\mathrm{L}_{\mathrm{R}}^{\dagger} = (\mathrm{Id} - \mathrm{Proj}_{\mathcal{P}})\mathrm{L}^{-1}\mathrm{R}^{*} : \mathcal{S}(\mathbb{R} \times \mathbb{S}^{d-1}) \to \mathcal{S}'(\mathbb{R}^d) \qquad \text{(Right-inverse of } \mathrm{L}_{\mathrm{R}}\text{)}$$

$$\mathrm{L}_{\mathrm{R}}^{\dagger *} = \mathrm{R}\mathrm{L}^{-1*}(\mathrm{Id} - \mathrm{Proj}_{\mathcal{P}'}) : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R} \times \mathbb{S}^{d-1}) \qquad \text{(Left-inverse of } \mathrm{L}_{\mathrm{R}}^{*}\text{)}$$

Theorem (Generalized impulse response of $\mathrm{L}_{\mathrm{R}}^{\dagger *}$) (Unser FoCM in press)
Let L be an admissible operator with a polynomial null space $\mathcal{P} = \mathcal{P}_{n_0}$ (possibly trivial) of degree n_0, a frequency profile $\widehat{L}_{\mathrm{rad}} : \mathbb{R} \to \mathbb{R}$, and $\rho_{\mathrm{rad}}(t) = \mathcal{F}^{-1}\{1/\widehat{L}_{\mathrm{rad}}\}(t)$. Then,

$$h\big(\boldsymbol{x}_0, (t, \boldsymbol{\xi})\big) \triangleq \mathrm{L}_{\mathrm{R}}^{\dagger *}\{\delta(\cdot - \boldsymbol{x}_0)\}(t, \boldsymbol{\xi})$$
$$= \rho_{\mathrm{rad}}(t - \boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{x}_0) - \sum_{n=0}^{n_0} \frac{(-\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{x}_0)^n}{n!}\big(\kappa_{\mathrm{rad}} * \partial^n \rho_{\mathrm{rad}}\big)(t)$$

with $h(\boldsymbol{x}_0, \cdot) \in C_0(\mathbb{R} \times \mathbb{S}^{d-1})$ and

$$\sup_{(\boldsymbol{x}_0, \boldsymbol{\xi}) \in \mathbb{R}^d \times \mathbb{S}^{d-1}} (1 + |\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{x}_0|)^{-n_0} \|h\big(\boldsymbol{x}_0; (\cdot, \boldsymbol{\xi})\big)\|_{L_q(\mathbb{R})} < \infty.$$

for any $q \in [2, \infty]$.

Kernel of stable right-inverse of "radonized" Laplacian

Continuity bound (non-trivial):

$$\big|\Delta_{\mathrm{R}}^{\dagger}\{w\}(\boldsymbol{x})\big| \le \|h(\boldsymbol{x}; \cdot)\|_{L_\infty} \|w\|_{\mathcal{M}} \le (1 + \|\boldsymbol{x}\|)\, \|w\|_{\mathcal{M}}$$

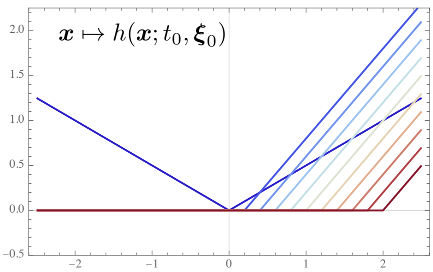
- Integral representation

$$\Delta_{\mathrm{R}}^{\dagger} : w \mapsto f(\boldsymbol{x}) = \int_{\mathbb{R}} \int_{\mathbb{S}^{d-1}} h(\boldsymbol{x}; t, \boldsymbol{\xi}) w(t, \boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}\mathrm{d}t \quad \text{where}$$

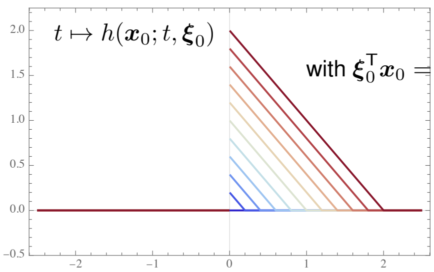
$$h(\boldsymbol{x}; t, \boldsymbol{\xi}) = \tfrac{1}{2}|\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{x} - t| - (\kappa_{\mathrm{rad}} * \tfrac{1}{2}|\cdot|)(t) + (\boldsymbol{\xi}^{\mathsf{T}}\boldsymbol{x})(\kappa_{\mathrm{rad}} * \tfrac{1}{2}\mathrm{sign})(t),$$

← Dictionary elements = extreme points of unit ball

where $\kappa_{\mathrm{rad}} \in \mathcal{S}_{\mathrm{even}}(\mathbb{R})$ is a normalized radial smoothing kernel.



along direction of the ridge; i.e., $\boldsymbol{x} = s\boldsymbol{\xi}_0$



with $\boldsymbol{\xi}_0^{\mathsf{T}}\boldsymbol{x}_0 = \mathrm{Const}$

(under simplifying assumption: $\kappa_{\mathrm{rad}} = \delta$)

Admissible (activation, operator) pairs

$\rho_{\mathrm{rad}}(t)$	$\tilde{\rho}_{\mathrm{rad}}(t)$	$\widehat{L}_{\mathrm{rad}}$ or $\widehat{\tilde{L}}_{\mathrm{rad}}$	n_0 (null space)
Exponential			
$\mathrm{e}^{-\lvert t\rvert}$	N/A	$1+\lvert\omega\rvert^2$	-1 (trivial)
	Classical sigmoids (Barron 1993)		
	$\frac{\tanh(\frac{t}{2})}{2} = \frac{1}{2} + \frac{1}{1+\mathrm{e}^{-t}}$	$\frac{\sinh(\pi\omega)}{\pi}$	0 (bias)
	$\frac{\arctan(t)}{\pi}$	$\omega \mathrm{e}^{\lvert\omega\rvert}$	0
Ridge splines	(of degree $n \in \mathbb{N}$)		
$\frac{1}{2}\lvert t\rvert$ (or ReLU)	$t \log\lvert t\rvert$	$\lvert\omega\rvert^2$	1 (affine)
$\propto t^{2n} \log\lvert t\rvert$	$\frac{\mathrm{sign}(t)\lvert t\rvert^{2n}}{(2n)!}$	$\lvert\omega\rvert^{2n+1}$	$2n \geq 2$ (even)
$\frac{1}{2}\frac{\lvert t\rvert^{2n+1}}{(2n+1)!}$	$\propto t^{2n+1} \log\lvert t\rvert$	$\lvert\omega\rvert^{2n+2}$	$2n+1 \geq 1$ (odd)
Fractional splines	(degree $\alpha \in \mathbb{R}^+ \backslash \mathbb{N}$)		
$\frac{\lvert t\rvert^\alpha \sin(\frac{\alpha\pi}{2})}{\pi\Gamma(\alpha)}$	$\frac{\mathrm{sign}(t)\lvert t\rvert^\alpha \cos(\frac{\alpha\pi}{2})}{\pi\Gamma(\alpha)}$	$\lvert\omega\rvert^{\alpha+1}$	$\lceil\alpha\rceil$

Table 1: Examples of admissible symmetric and anti-symmetric activation functions with their corresponding regularization operator. The anti-

- Anti-symmetric extension

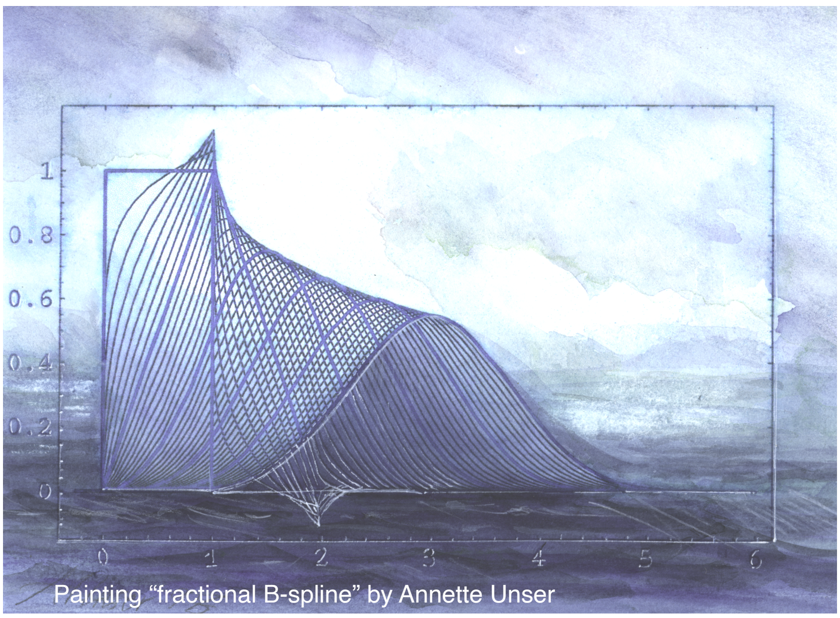
  Replace symmetric filtering operator $\mathrm{K}_{\mathrm{rad}}$ by anti-symmetric $\widetilde{\mathrm{K}}_{\mathrm{rad}} = \mathrm{H}_{\mathrm{rad}}\mathrm{K}_{\mathrm{rad}}$.

  $\mathrm{H}_{\mathrm{rad}}$: Hilbert transform

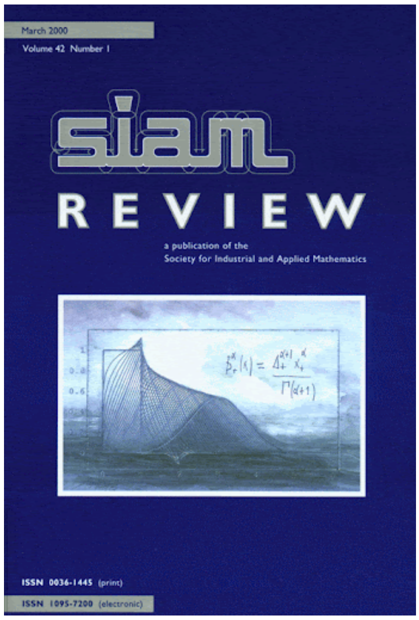
- Bottom line

  Choice of L (resp., L_{R}) and symmetry fixes activation $\sigma = \rho_{\mathrm{rad}}$ and vice versa

M. Unser, T. Blu, “Fractional Splines and Wavelets,” *SIAM Review*, March 2000.

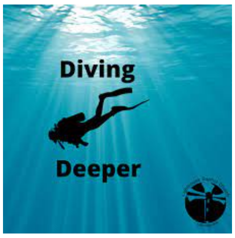


Painting “fractional B-spline” by Annette Unser



CONCLUSION: Return of the spline

- Foundations of **functional learning**
  - Functional optimization in Banach spaces (enabled by representer theorems)
  - **Hilbert spaces**: the tools of classical ML
  - **Non-reflexive Banach spaces**: for sparsity-promoting regularization (e.g., CS)
  - **Isotropy** + **Radon transform**: The key for obtaining pointwise nonlinearities
- Splines and machine learning
  - Traditional kernel methods are closely related to splines … and the same holds true for ReLU nets …
  - Sparsity-promoting regularization offer promising perspectives
  - Radon-domain regularization ⇒ *Unifies* **Shallow neural nets** and **RBF methods**



- Functional composition = **hierarchical splines**
  - Deep ReLU neural nets are **high-dimensional** piecewise-linear **splines**
  - **Free-form** activations with $\mathrm{TV}^{(2)}$-regularization ⇒ **Deep splines**

ACKNOWLEDGMENTS

Many thanks to (former) members of EPFL's Biomedical Imaging Group

- Prof. Rahul Parhi
- Dr. Shayan Aziznejad
- Dr. Julien Fageot
- Prof. John Paul Ward
- Dr. Thomas Debarre
- Dr. Mike McCann
- Dr. Harshit Gupta
- Prof. Kyong Jin
- Dr. Fangshu Yang
- Dr. Emrah Bostan
- Prof. Ulugbek Kamilov
- ....

CIBM Center for Biomedical Imaging

and collaborators ...

- Prof. Demetri Psaltis
- Prof. Marco Stampanoni
- Prof. Carlos-Oscar Sorzano
- Prof Luigi Ambrosio
- ….

erc 3 FunLearn

References http://bigwww.epfl.ch/

- Sparse adaptive splines
  - M. Unser, J. Fageot, J.P. Ward, "Splines Are Universal Solutions of Linear Inverse Problems with Generalized-TV Regularization," *SIAM Review*, vol. 59, No. 4, pp. 769-793, 2017.
  - T. Debarre, Q. Denoyelle, M. Unser, J. Fageot, "Sparsest Continuous Piecewise-Linear Representation of Data," *Journal of Computational and Applied Mathematics*, vol. 406, paper no. 114044, pp. 1-30, 2022.
- Representer theorems
  - M. Unser, "A Representer Theorem for Deep Neural Networks," *Journal of Machine Learning Research*, vol. 20, no. 110, pp. 1-30, Jul. 2019.
  - M. Unser, "A Unifying Representer Theorem for Inverse Problems and Machine Learning," *Foundations of Computational Mathematics*, vol. 21, pp. 941–960, 2021.
  - M. Unser, S. Aziznejad, "Convex optimization in sums of Banach spaces," *Applied and Computational Harmonic Analysis*, vol. 56, no. 1, pp. 1-25, 2022.
- Neural networks and the Radon transform
  - M. Unser, "Ridges, Neural Networks, and the Radon Transform", *Journal of Machine Learning Research*, vol. 24, no. 37, pp. 1-33, 2023.
  - S. Neumayer, M. Unser, "Explicit Representations for Banach Subspaces of Lizorkin Distributions," *Analysis and Applications* vol. 21, no. 5, pp. 1223–1250, September 2023. Preprint arXiv:2203.05312 [math.FA]
  - M. Unser, "Unifying Variational Formulation of Supervised Learning: From Kernel Methods to Neural Networks," *Foundations of Computational Mathematics* (in press). Preprint arXiv:2206.14625 [cs.LG]