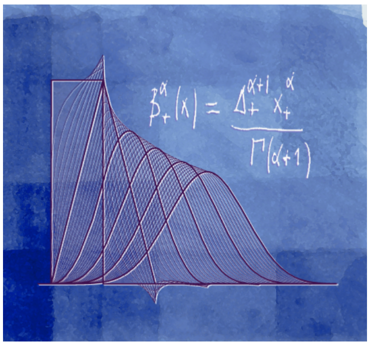
New representer theorems for inverse problems and machine learning

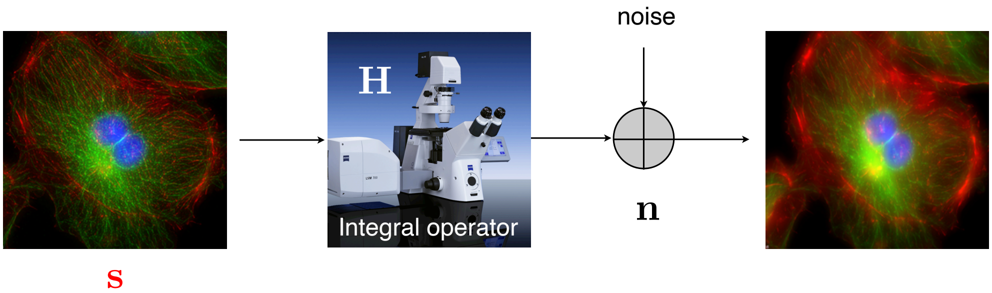
Michael Unser
Biomedical Imaging Group
EPFL, Lausanne, Switzerland



EPFL CIS – RIKEN AIP Seminar Series (virtual), March 2, 2022

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

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

Supervised learning as a (linear) inverse problem

but an infinite-dimensional one ...

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

- Introduce smoothness or **regularization** constraint (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

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

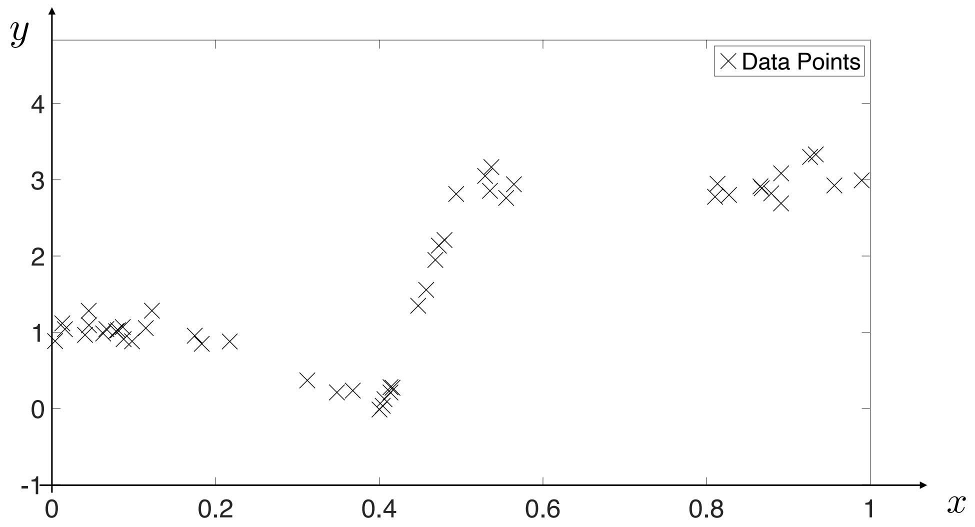
- Regularized least-squares fit (theory of RKHS)

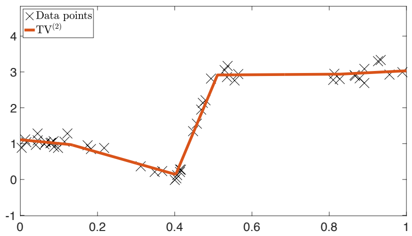
  $$f_{\mathrm{RKHS}} = \arg\min_{f \in \mathcal{H}} \left(\sum_{m=1}^{M} |y_m - f(\boldsymbol{x}_m)|^2 + \lambda R(f) \right) \quad \text{with} \quad R(f) = \|f\|_{\mathcal{H}}^2$$

  $\Rightarrow$ kernel estimator

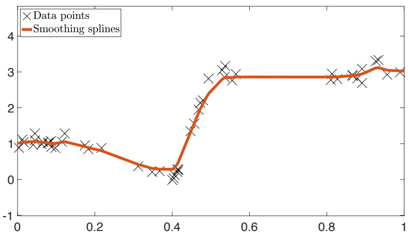
  (Wahba 1990; Schölkopf 2001)

Can your learn the map $y = f(x)$?

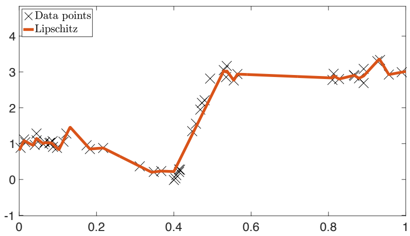




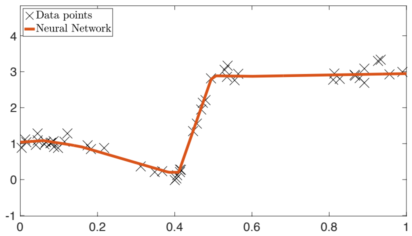
$R(f) = \|\mathrm{D}^2 f\|_{\mathcal{M}}$



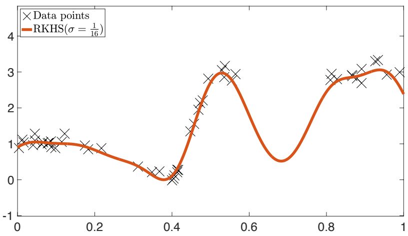
$R(f) = \|\mathrm{D}f\|_{L_2}^2$



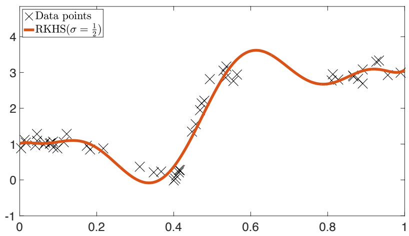
$R(f) = \|\mathrm{D}f\|_{L_\infty}$



DNN with (1,10,10,10,1)



$R(f) = \|f\|_{\mathcal{H}_{\mathrm{RKHS}_1}}^2$



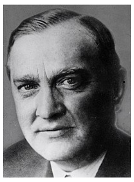
$R(f) = \|f\|_{\mathcal{H}_{\mathrm{RKHS}_2}}^2$

OUTLINE

- **Introduction** ✔
  - Learning as an inverse problem
  - Teaser: Search of the best learner
- **Foundations of functional learning**
  - Banach spaces and duality mappings
  - **Unifying representer theorem** NEW
- **From classical to modern regularization-based techniques**
  - Kernel methods
  - Smoothing splines
  - **Sparse adaptive splines**
- **Deep neural networks vs. deep splines**
  - Continuous piecewise linear (CPWL) functions / splines
  - **Representer theorem for deep neural networks**

Swiss National Science Foundation

General notion of Banach space



Stefan Banach (1892-1945)

Normed space: vector space $\mathcal{X}$ equipped with a norm $\|\cdot\|_{\mathcal{X}}$

Convergent sequence of functions (φ_i) in $\mathcal{X}$:

$$\lim_i \varphi_i = \varphi; \quad \text{i.e., } \lim_i \|\varphi - \varphi_i\|_{\mathcal{X}} = 0$$

> **Definition**
> A Banach space is a **complete normed** space $\mathcal{X}$;
> that is, such that $\lim_i \varphi_i = \varphi \in \mathcal{X}$ for any convergent sequence (φ_i) in $\mathcal{X}$.

- Generality of the concept
  - Linear space of vectors $\boldsymbol{u} = (u_1, \dots, u_N) \in \mathbb{R}^N$
  - Linear space of functions $u : \mathbb{R}^d \to \mathbb{R}$
  - Linear space of vector-valued functions $\boldsymbol{u} = (u_1, \dots, u_N) : \mathbb{R}^d \to \mathbb{R}^N$
  - Space of linear functional $u : \mathcal{X} \to \mathbb{R}$
  - Linear space $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ of bounded operators $\mathrm{U} : \mathcal{X} \to \mathcal{Y}$

Dual of a Banach space

Dual of the Banach space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$:
$\mathcal{X}'$ = space of linear functionals $g : f \mapsto \langle g, f\rangle \triangleq g(f) \in \mathbb{R}$ that are continuous on $\mathcal{X}$

$\mathcal{X}'$ is a Banach space equipped with the **dual norm**:

$$\|g\|_{\mathcal{X}'} = \sup_{f \in \mathcal{X}\setminus\{0\}} \left(\frac{\langle g, f\rangle}{\|f\|_{\mathcal{X}}} \right)$$

$$\Rightarrow \quad \|g\|_{\mathcal{X}'} \geq \frac{|\langle g, f\rangle|}{\|f\|_{\mathcal{X}}}, \quad f \neq 0$$

- Generic duality bound

> For any $f \in \mathcal{X}, g \in \mathcal{X}'$: $\quad |\langle g, f\rangle| \leq \|g\|_{\mathcal{X}'} \|f\|_{\mathcal{X}}$

- Duals of L_p spaces: $\quad \left(L_p(\mathbb{R}^d)\right)' = L_{p'}(\mathbb{R}^d)$ with $\frac{1}{p} + \frac{1}{p'} = 1$ for $p \in (1, \infty)$

Hölder inequality: $\quad |\langle f, \varphi\rangle| \leq \int_{\mathbb{R}^d} |f(\boldsymbol{r})\varphi(\boldsymbol{r})| \, \mathrm{d}\boldsymbol{r} \leq \|f\|_{L_p} \|\varphi\|_{L_{p'}}$

Riesz conjugate for Hilbert spaces

- Duality bound for Hilbert spaces (equivalent to Cauchy-Schwarz inequality)

Frigyes Riesz (1880-1956)

> For all $(u, v) \in \mathcal{H} \times \mathcal{H}'$: $\quad |\langle u, v\rangle| \leq \|u\|_{\mathcal{H}} \, \|v\|_{\mathcal{H}'}$

- Definition

The **Riesz conjugate** of $u \in \mathcal{H}$ is the unique element $u^* \in \mathcal{H}'$ such that

$$\langle u, u^* \rangle = \langle u, u \rangle_{\mathcal{H}} = \|u\|_{\mathcal{H}}^2 = \|u\|_{\mathcal{H}} \, \|u^*\|_{\mathcal{H}'}$$

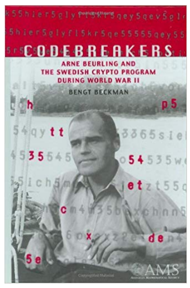
(sharp duality bound)

\+

(isometry)

- Properties
  - Norm preservation: $\|u\|_{\mathcal{H}} = \|u^*\|_{\mathcal{H}'}$
  - $u^* = \mathrm{R}^{-1}\{u\}$ (inverse Riesz map)
  - Invertibility: $u = (u^*)^* = \mathrm{R}\{u^*\}$ $\qquad (\mathcal{H}')' = \mathcal{H}$ (reflexivity)
  - Linearity: $(u_1 + u_2)^* = u_1^* + u_2^*$

Generalization: Duality mapping



Arne Beurling (1905-1986)

Definition
Let $(\mathcal{X}, \mathcal{X}')$ be a dual pair of Banach spaces. Then, the elements $f^* \in \mathcal{X}'$ and $f \in \mathcal{X}$ form a **conjugate pair** if

- $\|f^*\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$ (norm preservation), and
- $\langle f^*, f\rangle_{\mathcal{X}' \times \mathcal{X}} = \|f^*\|_{\mathcal{X}'} \|f\|_{\mathcal{X}}$ (sharp duality bound).

For any given $f \in \mathcal{X}$, the set of admissible conjugates defines the **duality mapping**

$$J(f) = \{f^* \in \mathcal{X}' : \|f^*\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}} \text{ and } \langle f^*, f\rangle_{\mathcal{X}' \times \mathcal{X}} = \|f^*\|_{\mathcal{X}'} \|f\|_{\mathcal{X}}\},$$

which is a non-empty subset of $\mathcal{X}'$. Whenever the duality mapping is single-valued (for instance, when $\mathcal{X}'$ is strictly convex), one also defines the duality operator $\mathrm{J}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}'$, which is such that $f^* = \mathrm{J}_{\mathcal{X}}(f)$.

(Beurling-Livingston, 1962)

Properties of duality mapping

Theorem
Let $(\mathcal{X}, \mathcal{X}')$ be a dual pair of Banach spaces. Then, the following holds:

1. Every $f \in \mathcal{X}$ admits at least one conjugate $f^* \in \mathcal{X}'$.
2. For every $f \in \mathcal{X}$, the set $J(f)$ is convex and weak-$*$ closed in $\mathcal{X}'$.
3. The duality mapping is **single-valued** if $\mathcal{X}'$ is **strictly convex**; the latter condition is also necessary if $\mathcal{X}$ is reflexive.

$\mathcal{X}$ is ***strictly convex*** if, for all $f_1, f_2 \in \mathcal{X}$ such that $\|f_1\|_{\mathcal{X}} = \|f_2\|_{\mathcal{X}} = 1$ and $f_1 \neq f_2$, one has $\|\lambda f_1 + (1-\lambda) f_2\|_{\mathcal{X}} < 1$ for any $\lambda \in (0,1)$.

$\mathcal{X}$ is ***reflexive*** if $\mathcal{X}'' = \mathcal{X}$.

Mother of all representer theorems

$$\arg\min_{f \in \mathcal{X}'} E\big(\boldsymbol{y}, \boldsymbol{\nu}(f)\big) + \psi\left(\|f\|_{\mathcal{X}'}\right)$$


Lausanne, Christmas 2018

Mathematical assumptions:

- $(\mathcal{X}, \mathcal{X}')$ is a dual pair of Banach spaces.
- $\mathcal{N}_{\boldsymbol{\nu}} = \mathrm{span}\{\nu_m\}_{m=1}^M \subset \mathcal{X}$ with the ν_m being linearly independent.
- $\boldsymbol{\nu} : \mathcal{X}' \to \mathbb{R}^M : f \mapsto \big(\langle \nu_1, f\rangle, \dots, \langle \nu_M, f\rangle\big)$ is the **linear measurement operator** (it is weak* **continuous** on $\mathcal{X}'$ because $\nu_1, \dots, \nu_M \in \mathcal{X}$).
- $E : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}^+$ is a strictly-convex loss functional.
- $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ is some arbitrary strictly-increasing convex function.

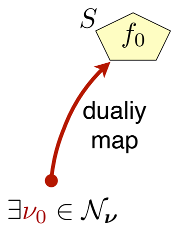
General representer theorem

Theorem

For any fixed $\boldsymbol{y} \in \mathbb{R}^M$, the solution set of the **generic** optimization problem

$$S = \arg \min_{f \in \mathcal{X}'} E\big(\boldsymbol{y}, \boldsymbol{\nu}(f)\big) + \psi\left(\|f\|_{\mathcal{X}'}\right)$$

is **non-empty, convex** and weak*-compact, and all solutions $f_0 \in S \subset \mathcal{X}'$ are $(\mathcal{X}', \mathcal{X})$-conjugate of **a common** $\nu_0 \in \mathcal{N}_{\boldsymbol{\nu}} = \mathrm{span}\{\nu_m\}_{m=1}^M \subset \mathcal{X}$.



The parametric form of the solution depends on the space type.

1) If $\mathcal{X}'$ is a **Hilbert space** and ψ is strictly convex, then the solution is unique and it admits a **linear expansion** with coefficients $(a_m) \in \mathbb{R}^M$

$$f_0 = \sum_{m=1}^{M} a_m \varphi_m,$$

where $\varphi_m = \mathrm{J}_{\mathcal{X}}\{\nu_m\} \in \mathcal{X}'$ with $\mathrm{J}_{\mathcal{X}}$ the Riesz map $\mathcal{X} \to \mathcal{X}'$.

(Unser, FoCM 2021)

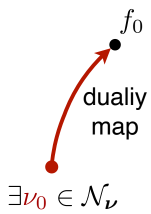
General representer theorem (Cont'd)

2) If $\mathcal{X}'$ is a **strictly convex Banach space** and ψ is strictly convex, then the solution is unique and it admits the **representation** with $(a_m) \in \mathbb{R}^M$

$$f_0 = \mathrm{J}_{\mathcal{X}} \left\{ \sum_{m=1}^{M} a_m \nu_m \right\},$$

where $\mathrm{J}_{\mathcal{X}}$ is the (nonlinear) duality operator $\mathcal{X} \to \mathcal{X}'$.

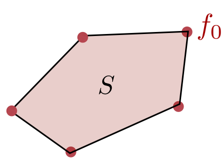
(Unser, FoCM 2021)



3) Otherwise, when $\mathcal{X}'$ is **not strictly convex**, the solution set S is the convex hull of its **extreme points**, which can all be expressed as

$$f_0 = \sum_{k=1}^{K_0} c_k e_k,$$

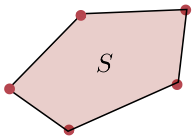
for some $K_0 \leq M$, $c_1, \dots, c_{K_0} \in \mathbb{R}$, where $e_1, \dots, e_{K_0} \in \mathcal{X}'$ are some extreme points of the unit ball $B_{\mathcal{X}'} = \{x \in \mathcal{X} : \|x\|_{\mathcal{X}'} \leq 1\}$.



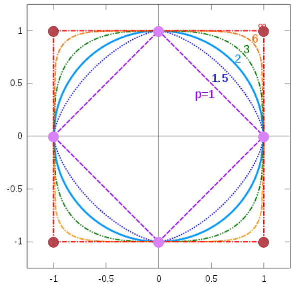
(Boyer-Chambolle-De Castro-Duval-De Gournay-Weiss, arXiv:1806.09810, 2019)

Extreme points

- Definition



Let S be a convex set. Then, the point $x \in S$ is **extreme**
if it cannot be expressed as a (non-trivial) convex combination of any other points in S.



- Extreme points of unit ball in $\ell_p(\mathbb{Z})$
  - $\ell_\infty(\mathbb{Z}): \quad e_k[n] = \pm 1$
  - $\ell_1(\mathbb{Z}): \quad e_k = \pm\delta[\cdot - n_k]$ (Kronecker impulse) $\Rightarrow$ sparse !!!
  - $\ell_p(\mathbb{Z})$ with $p \in (1, \infty)$: $\quad e_k = u/\|u\|_{\ell_p}$ for any $u \in \ell_p(\mathbb{Z})$

Definition of ***strictly convex*****ity** of a Banach space: all boundary points are extreme !!!

OUTLINE

- Introduction ✔
- Foundations of functional learning ✔
- **From classical to modern regularization-based techniques**
  - Learning in RKHS
  - Kernel methods of ML
  - Smoothing splines
  - **Sparse kernel methods**
  - **Sparse adaptive splines**
  - **Lipchitz splines**

- Deep neural networks vs. deep splines

1. Learning in reproducing kernel Hilbert space

Definition
A Hilbert space $\mathcal{H}$ of functions on $\mathbb{R}^d$ is called a **reproducing kernel Hilbert space** (RKHS) if $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{H}'$ for any $\boldsymbol{x}_0 \in \mathbb{R}^d$. The corresponding unique **Hilbert conjugate** $h(\cdot, \boldsymbol{x}_0) = \left(\delta(\cdot - \boldsymbol{x}_0)\right)^* \in \mathcal{H}$ when indexed by $\boldsymbol{x}_0$ is called the **reproducing kernel** of $\mathcal{H}$.

- Learning problem

Given the data $\left(\boldsymbol{x}_m, y_m\right)_{m=1}^M$ with $\boldsymbol{x}_m \in \mathbb{R}^d$, find the function $f_0 : \mathbb{R}^d \to \mathbb{R}$ s.t.

$$f_0 = \arg\min_{f \in \mathcal{H}} \left(\sum_{m=1}^{M} E_m\big(y_m, f(\boldsymbol{x}_m)\big) + \psi(\|f\|_{\mathcal{H}}) \right)$$

  - $E_m : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ (strictly convex)
  - $\psi : \mathbb{R} \to \mathbb{R}^+$ (strictly increasing and convex)

Learning in RKHS (Cont'd)

- Special case of general representer theorem
  - $\mathcal{X} = \mathcal{H}', \ \mathcal{X}' = \mathcal{H}'' = \mathcal{H}$ (all Hilbert spaces are reflexive)
  - $\nu_m = \delta(\cdot - \boldsymbol{x}_m)$ (Dirac sampling functionals)
  - Additive loss: $E(\boldsymbol{y}, \boldsymbol{z}) = \sum_{m=1}^{M} E_m\big(y_m, z_m\big)$

specific of ML

- Key observation

Reproducing kernel = Schwartz kernel of **Riesz map**

$$\mathrm{R} = \mathrm{J}_{\mathcal{H}'} : \mathcal{H}' \to \mathcal{H} : \nu \mapsto \int_{\mathbb{R}^d} h(\cdot, \boldsymbol{y})\nu(\boldsymbol{y})\mathrm{d}\boldsymbol{y} \quad \Rightarrow \quad \varphi_m = \mathrm{J}_{\mathcal{H}'}\{\delta(\cdot - \boldsymbol{x}_m)\} = h(\cdot, \boldsymbol{x}_m)$$

- Implied form of unique solution = linear kernel expansion

$$f_0(\boldsymbol{x}) = \sum_{m=1}^{M} a_m \varphi_m(\boldsymbol{x}) = \sum_{m=1}^{M} a_m h(\boldsymbol{x}, \boldsymbol{x}_m)$$

(Schölkopf representer theorem, 2001)

2. Regularization with a LSI operator = kernel methods of ML

- Quadratic Tikhonov regularization functional

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

(Poggio-Girosi 1990)

L: Linear shift-invariant (LSI), **invertible regularization operator**

$\widehat{L}(\boldsymbol{\omega})$: **frequency response** of L

Hilbertian isometries

$$\mathcal{H}' \underset{\mathrm{L}^*}{\overset{\mathrm{L}^{-1*}}{\rightleftarrows}} L_2(\mathbb{R}^d) \underset{\mathrm{L}}{\overset{\mathrm{L}^{-1}}{\rightleftarrows}} \mathcal{H}$$

- Key observation

Reproducing kernel = Impulse response of $\mathrm{L}^{-1}\mathrm{L}^{-1*} = (\mathrm{L}^*\mathrm{L})^{-1}$

$\nu^* = \mathrm{J}_{\mathcal{H}'}\{\nu\} = h * \nu$ where $h = \mathcal{F}^{-1}\left\{\frac{1}{|\widehat{L}(\boldsymbol{\omega})|^2}\right\} \in L_1(\mathbb{R}^d)$

- Parametric form of solution = expansion of **kernels** centered on data points

$$f_0(\boldsymbol{x}) = \sum_{m=1}^{M} a_m \mathrm{J}_{\mathcal{H}'}\{\delta(\cdot - \boldsymbol{x}_m)\}(\boldsymbol{x}) = \sum_{m=1}^{M} a_m h(\boldsymbol{x} - \boldsymbol{x}_m)$$

3. Smoothing splines

$$f_0 = \arg \min_{f:\, \mathbb{R}\to\mathbb{R}} \left(\sum_{m=1}^{M} |f(x_m) - y_m|^2 + \lambda \int_{\mathbb{R}} \left| \frac{\mathrm{d}f(x)}{\mathrm{d}x} \right|^2 \mathrm{d}x \right)$$

(Schoenberg 1964; de Boor 1966)

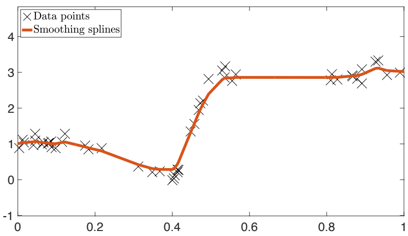
- Smoothness regularization (spline semi-norm)

$R(f) = \|\mathrm{D}f\|_{L_2}^2$ with $\mathrm{D} = \frac{\mathrm{d}}{\mathrm{d}x}$; **Null space** : $\mathcal{N}_{\mathrm{D}} = \{p(x) = a_0 : a_0 \in \mathbb{R}\}$

- Direct-sum RKHS topology: $L_{2,\mathrm{D}}(\mathbb{R}) = \mathcal{H}_{\mathrm{D}} \oplus \mathcal{N}_{\mathrm{D}}$

D has a unique inverse only if one factors out the null space

Impulse response of $(\mathrm{D}^*\mathrm{D})^{-1}$: $h(x) = \mathcal{F}^{-1}\left\{\frac{1}{|\omega|^2}\right\}(x) = \frac{1}{2}|x|$



- Solution = linear spline with knots at $x_1, \dots, x_M$

$$f_0(x) = a_0 + \sum_{m=1}^{M} a_m |x - x_m|$$

4. Sparse kernel expansions

- Sparsity-promoting regularization functional

$$R(f) = \cancel{\|\mathrm{L}f\|_{L_1} = \int_{\mathbb{R}^N} |\mathrm{L}f(\boldsymbol{x})|\mathrm{d}\boldsymbol{x}}$$

L: Linear shift-invariant (LSI), **invertible regularization operator**

$\widehat{L}(\boldsymbol{\omega})$: **frequency response** of L

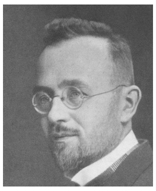
Banach isometry

$$L_1(\mathbb{R}^d) \underset{\mathrm{L}}{\overset{\mathrm{L}^{-1}}{\rightleftarrows}} L_{1,\mathrm{L}}(\mathbb{R}^d)$$

Theoretical roadblock: The general representer theorem does not apply because **there exists no predual space** $\mathcal{X}$ such that $L_1(\mathbb{R}^d) = \mathcal{X}'$.

The optimization problem is ill-defined and does not admit a solution !

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 $\boldsymbol{\tau}_k \in \mathbb{R}^d$

4. Sparse kernel expansions (2nd attempt)

- Sparsity-promoting regularization functional

$$R(f) = \|\mathrm{L}f\|_{\mathcal{M}} = \sup_{\varphi \in C_0(\mathbb{R}^d): \|\varphi\|_{L_\infty} \le 1} \langle \mathrm{L}f, \varphi \rangle$$

L: Linear shift-invariant (LSI), **invertible regularization operator**

$\widehat{L}(\boldsymbol{\omega})$: **frequency response** of L

Impulse response of L^{-1}: $h = \mathcal{F}^{-1}\left\{\frac{1}{\widehat{L}(\boldsymbol{\omega})}\right\} \in L_1(\mathbb{R}^d)$

Banach isometry

$$\mathcal{M}(\mathbb{R}^d) \underset{\mathrm{L}}{\overset{\mathrm{L}^{-1}}{\rightleftarrows}} \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$$

Extreme points: $e_k = \mathrm{L}^{-1}\{\delta(\cdot - \boldsymbol{\tau}_k)\}$

(Aziznejad-U., SIAM 2021)

Corollary (3rd case of representer theorem)

The extreme points f_0 of $S = \arg \min_{f \in \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)} \left(\sum_{m=1}^{M} E_m\big(y_m, f(\boldsymbol{x}_m)\big) + \lambda \|\mathrm{L}f\|_{\mathcal{M}} \right)$ can all be expressed as

$$f_0(\boldsymbol{x}) = \sum_{k=1}^{K_0} a_k h(\boldsymbol{x} - \boldsymbol{\tau}_k)$$

for some $K_0 \le M$, $\boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_{K_0} \in \mathbb{R}^d$ and $\boldsymbol{a} = (a_k) \in \mathbb{R}^{K_0}$. Moreover, $\|\mathrm{L}f_0\|_{\mathcal{M}} = \sum_{k=1}^{K_0} |a_k| = \|\boldsymbol{a}\|_{\ell_1}$.

5. Sparse adaptive spline

$$f_0 = \arg \min_{f \in \mathcal{M}_{\mathrm{D}^2}(\mathbb{R})} \left(\sum_{m=1}^{M} |f(x_m) - y_m|^2 + \lambda \|\mathrm{D}^2 f\|_{\mathcal{M}} \right)$$

(Mammen 1997; Unser 2017)

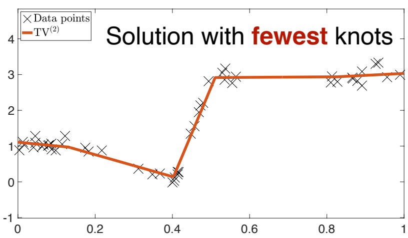
- Sparsity-promoting regularization

$R(f) = \|\mathrm{D}^2 f\|_{\mathcal{M}}$ **Null space** : $\mathcal{N}_{\mathrm{D}^2} = \{p(x) = b_0 + b_1 x : b_0, b_1 \in \mathbb{R}\}$

- Direct-sum Banach topology: $\mathcal{M}_{\mathrm{D}^2}(\mathbb{R}) = \mathcal{U}_{\mathrm{D}^2} \oplus \mathcal{N}_{\mathrm{D}^2}$

D^2 has a unique invertise only if one factors out the null space

Impulse response of D^{-2} (two-fold integrator): $h(x) = (x)_+ = \mathrm{ReLU}(x)$



(Debarre arXiv 2020)

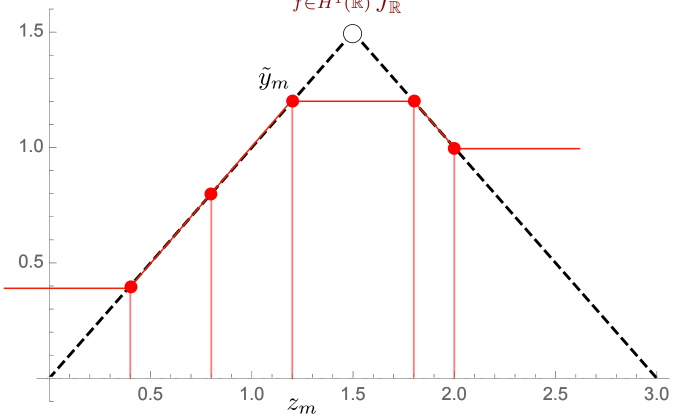
- Solution = **linear spline** with (few) **adaptive** knots at $\tau_1, \dots, \tau_{K_0}$

$$f_0(x) = b_0 + b_1 x + \sum_{k=1}^{K_0} a_k (x - \tau_k)_+$$

Comparison of linear interpolators

$$\arg\min_{f\in H^1(\mathbb{R})}\int_{\mathbb{R}}|\mathrm{D}f(x)|^2\mathrm{d}x \quad \text{s.t.} \quad f(x_m)=y_m,\ m=1,\dots,M$$

(de Boor 1966)



$$\arg\min_{f\in \mathrm{BV}^{(2)}(\mathbb{R})}\|\mathrm{D}^2 f\|_{\mathcal{M}} \quad \text{s.t.} \quad f(x_m)=y_m,\ \ m=1,\dots,M$$

(Unser JMLR 2019; Lemma 2)

6. Lipschitz splines

$$f_0=\arg\min_{f\in W^1_\infty(\mathbb{R})}\left(\sum_{m=1}^{M}|f(x_m)-y_m|^2+\lambda\|\mathrm{D}f\|_{L_\infty}\right)$$

- Lipschitz boundedness constraint

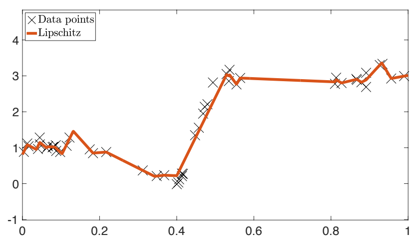
  $R(f)=\|\mathrm{D}f\|_{L_\infty}$ **Null space** : $\mathcal{N}_{\mathrm{D}}=\{p(x)=b_0 : b_0\in\mathbb{R}\}$

  Extreme points of unit ball in $L_\infty(\mathbb{R})$: e_k such that $e_k(x)=\pm 1$

- Direct-sum Banach topology: $W^1_\infty(\mathbb{R})=\mathcal{U}_{\mathrm{D}}\oplus\mathcal{N}_{\mathrm{D}}$

  D has a unique inverse only if one factors out the null space

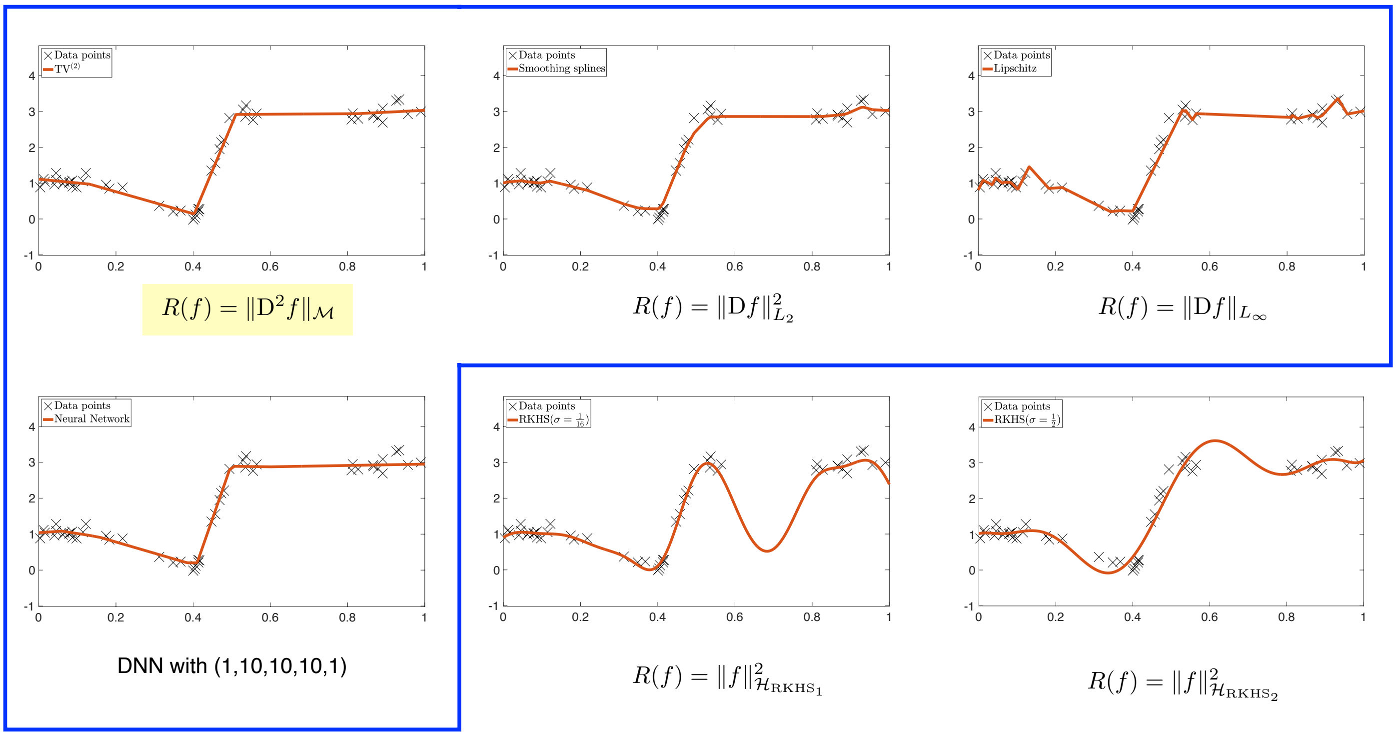
  $u_k=\mathrm{D}^{-1}e_k(x)=\int_{-\infty}^{x}e_k(t)\mathrm{d}t+C_k$: linear spline with binary slope (± 1)



- Solution = **linear spline** with with many oscillations (non-unique)

$$f_0(x)=b_0+\sum_{k=1}^{K_0}a_k u_k(x)$$

(Aziznejad et al., ArXiv 2022)

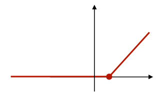


OUTLINE

- Introduction ✔
- Foundations of functional learning ✔
- From classical to modern regularization-based techniques ✔
- **Deep neural networks vs. deep splines**
  - Background
  - Continuous piecewise linear (CPWL) functions / splines
  - Variational formulation of shallow nets
  - Representer theorem for deep neural networks

Deep neural networks and splines

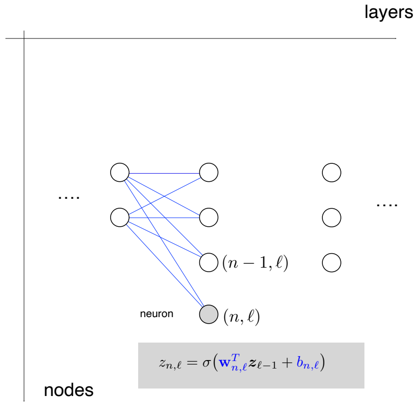
$\mathrm{ReLU}(x; b) = (x - b)_+$



- Preferred choice of activation function: ReLU
  - ReLU works nicely with dropout / ℓ_1-regularization — (Glorot *ICAIS* 2011)
  - Networks with hidden ReLU are easier to train
  - State-of-the-art performance — (LeCun-Bengio-Hinton *Nature* 2015)
- Deep nets as Continuous PieceWise-Linear maps
  - ReLU $\Rightarrow$ CPWL — (Montufar *NIPS* 2014)
  - CPWL $\Rightarrow$ Deep ReLU network — (Strang *SIAM News 2018*)
- Deep ReLU nets = hierarchical splines
  - ReLU is a piecewise-linear spline — (Poggio-Rosasco 2015)

Feedforward deep neural network

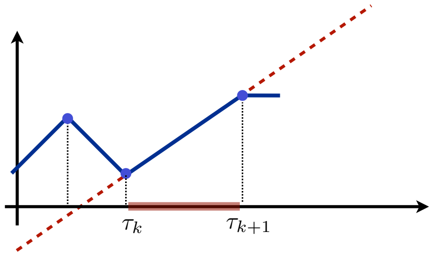
- Layers: $\ell = 1, \dots, L$
- Deep structure descriptor: $(N_0, N_1, \cdots, N_L)$
- Neuron or node index: $(n, \ell), \quad n = 1, \cdots, N_\ell$
- Activation function: $\sigma : \mathbb{R} \to \mathbb{R}$ (ReLU)
- Linear step: $\mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$
  $\boldsymbol{f}_\ell : \boldsymbol{x} \mapsto \boldsymbol{f}_\ell(\boldsymbol{x}) = \mathbf{W}_\ell \boldsymbol{x} + \mathbf{b}_\ell$
- Nonlinear step: $\mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$
  $\boldsymbol{\sigma}_\ell : \boldsymbol{x} \mapsto \boldsymbol{\sigma}_\ell(\boldsymbol{x}) = \big(\sigma(x_1), \dots, \sigma(x_{N_\ell})\big)$



Learned

$$\boxed{\mathrm{f}_{\text{deep}}(\boldsymbol{x}) = (\boldsymbol{\sigma}_L \circ \boldsymbol{f}_L \circ \boldsymbol{\sigma}_{L-1} \circ \cdots \circ \boldsymbol{\sigma}_2 \circ \boldsymbol{f}_2 \circ \boldsymbol{\sigma}_1 \circ \boldsymbol{f}_1)(\boldsymbol{x})}$$

Continuous-PieceWise Linear (CPWL) functions



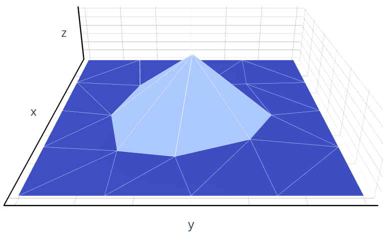
- 1D: Non-uniform spline de degree 1

Partition: $\mathbb{R} = \bigcup_{k=0}^{K} P_k$ with $P_k = [\tau_k, \tau_{k+1})$, $\tau_0 = -\infty < \tau_1 < \cdots < \tau_K < \tau_{K+1} = +\infty$.

The function $f_{\mathrm{spline}} : \mathbb{R} \to \mathbb{R}$ is a piecewise-linear spline with knots $\tau_1, \ldots, \tau_K$ if

- (i) for $x \in P_k$: $f_{\mathrm{spline}}(x) = f_k(x) \triangleq a_k x + b_k$ with $(a_k, b_k) \in \mathbb{R}^2$, $k = 0, \ldots, K$
- (ii) f_{spline} is continuous $\mathbb{R} \to \mathbb{R}$
- $f_{\mathrm{spline}}(x) = \tilde{b}_0 + \tilde{b}_1 x + \sum_{k=1}^{K} \tilde{a}_k (x - \tau_k)_+$ with $\tilde{b}_0, \tilde{b}_1 \in \mathbb{R}$, $(\tilde{a}_k) \in \mathbb{R}^K$.

CPWL functions in high dimensions



- Multidimensional generalization

Partition of domain into a finite number of non-overlapping **convex polytopes**; i.e.,

$\mathbb{R}^N = \bigcup_{k=1}^{K} P_k$ with $\mu(P_{k_1} \cap P_{k_2}) = 0$ for all $k_1 \neq k_2$

The function $f_{\mathrm{CPWL}} : \mathbb{R}^N \to \mathbb{R}$ is **continuous piecewise-linear** with partition $P_1, \ldots, P_K$

- (i) for $\boldsymbol{x} \in P_k$: $f_{\mathrm{CPWL}}(\boldsymbol{x}) = f_k(\boldsymbol{x}) \triangleq \mathbf{a}_k^T \boldsymbol{x} + b_k$ with $\mathbf{a}_k \in \mathbb{R}^N, b_k \in \mathbb{R}$, $k = 1, \ldots, K$
- (ii) f_{CPWL} is continuous $\mathbb{R}^N \to \mathbb{R}$

The vector-valued function $\mathbf{f}_{\mathrm{CPWL}} = (f_1, \ldots, f_M) : \mathbb{R}^N \to \mathbb{R}^M$ is a CPWL
if each component function $f_m : \mathbb{R}^N \to \mathbb{R}$ is CPWL.

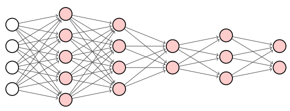
Algebra of CPWL functions

- any linear combination of (vector-valued) CPWL functions $\mathbb{R}^N \to \mathbb{R}^{N'}$ is CPWL, and,
- the composition $\mathbf{f}_2 \circ \mathbf{f}_1$ of any two CPWL functions with compatible domain and range—i.e., $\mathbf{f}_2 : \mathbb{R}^{N_1} \to \mathbb{R}^{N_2}$ and $\mathbf{f}_1 : \mathbb{R}^{N_0} \to \mathbb{R}^{N_1}$—is CPWL $\mathbb{R}^{N_0} \to \mathbb{R}^{N_2}$.

Sketch of proof: The continuity property is preserved through composition. The composition of two affine transforms is an affine transform, including the scenari where the domain is partitioned.

- The max (resp. min) pooling of two (or more) CPWL functions is CPWL.

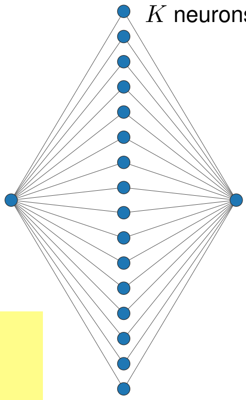
Implication for deep ReLU neural networks



$$\mathbf{f}_{\text{deep}}(\boldsymbol{x}) = (\boldsymbol{\sigma}_L \circ \boldsymbol{f}_L \circ \boldsymbol{\sigma}_{L-1} \circ \cdots \circ \boldsymbol{\sigma}_2 \circ \boldsymbol{f}_2 \circ \boldsymbol{\sigma}_1 \circ \boldsymbol{f}_1)(\boldsymbol{x})$$

- Each scalar neuron activation, $\sigma_{n,\ell}(x) = \mathrm{ReLU}(x)$, is CPWL.
- Each layer function $\boldsymbol{\sigma}_\ell \circ \boldsymbol{f}_\ell(\boldsymbol{x}) = (\mathbf{W}_\ell \boldsymbol{x} + \mathbf{b}_\ell)_+$ is CPWL
- The whole feedforward network $\mathbf{f}_{\text{deep}} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ is CPWL
- This holds true as well for deep architectures that involve Max pooling for dimension reduction
- The CPWL also remains valid for more complicated neuronal responses as long as they are CPWL; that is, **linear splines**.

Limit behaviour of univariate shallow ReLU neural nets



- Shallow univariate ReLU neural network with skip connection

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

- Standard training with weight decay

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

Theorem

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

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

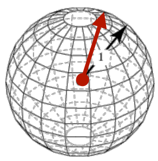
Arguments for the proof:

(Savarese 2019; Parhi-Nowak 2020)

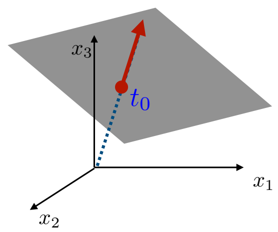
- Scale invariance of ReLU architecture: For any $\gamma > 0$, the map $(v_k, w_k) \mapsto (\gamma v_k, w_k/\gamma)$ does not affect $f_{\boldsymbol{\theta}}$.
- At the optimum of (NN-1), $|w_k| = |v_k|$, for $k = 1, \dots, K$ and $\mathrm{TV}^{(2)}(f_{\boldsymbol{\theta}}) = \sum_{k=1}^{K} |a_k|$ with $a_k = v_k |w_k|$.

The Radon transform and the FBP algorithm

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



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 $f \in L_1(\mathbb{R}^d)$

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

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

$$f = \mathrm{R}^* \mathrm{K}_{\mathrm{rad}}\{g\}$$

  - $\mathrm{K}_{\mathrm{rad}}$: "radial" filtering in Radon space along the variable t.
    Frequency response: $\widehat{K}_{\mathrm{rad}}(\omega) \propto |\omega|^{d-1}$
  - R^*: backprojection operator (the adjoint of R)

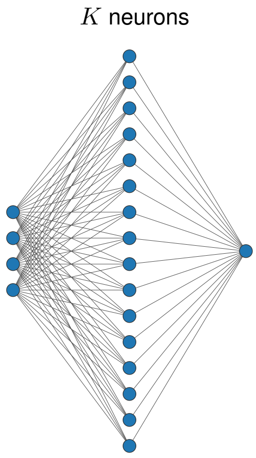
Limit behaviour of multivariate 2-layer ReLU neural nets

- 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 $\mathbf{W} = [\boldsymbol{w}_1 \ \dots \boldsymbol{w}_K]$

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



Theorem

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

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

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

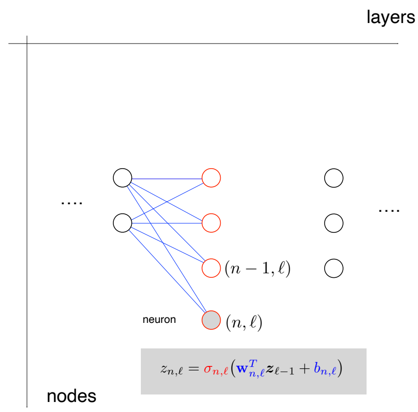
(Ongie et al. 2020; Parhi-Nowak 2021)

- $\mathcal{M}_{\text{Rad}}(\mathbb{R} \times \mathbb{S}^{d-1})$: space of bounded Radon-compatible measures
  $\mathcal{M}_{\text{Rad}} \subset \mathcal{S}'_{\text{Rad}} = \mathrm{K}_{\text{rad}}\mathrm{R}\big(\mathcal{S}'(\mathbb{R}^d)\big)$
- $\mathcal{M}_{\Delta_{\mathrm{R}}}(\mathbb{R}^d)$ = Banach space that is isometrically isomorphic to $\mathcal{M}_{\text{Rad}} \times \{c_0 + \boldsymbol{c}_1^{\mathsf{T}}\boldsymbol{x}\}$
- Regularization operator $\quad \Delta_{\mathrm{R}} = \mathrm{K}_{\text{rad}}\mathrm{R}\Delta : \mathcal{M}_{\Delta_{\mathrm{R}}}(\mathbb{R}^d) \to \mathcal{M}_{\text{Rad}}$

Refinement: free-form activation functions

- Layers: $\ell = 1, \dots, L$
- Deep structure descriptor: $(N_0, N_1, \cdots, N_L)$
- Neuron or node index: $(n, \ell), \ \ n = 1, \cdots, N_\ell$
- Activation function: $\sigma : \mathbb{R} \to \mathbb{R}$ (ReLU)

- Linear step: $\mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$
  $\boldsymbol{f}_\ell : \boldsymbol{x} \mapsto \boldsymbol{f}_\ell(\boldsymbol{x}) = \mathbf{W}_\ell\boldsymbol{x} + \mathbf{b}_\ell$
- Nonlinear step: $\mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$
  $\boldsymbol{\sigma}_\ell : \boldsymbol{x} \mapsto \boldsymbol{\sigma}_\ell(\boldsymbol{x}) = \big(\sigma_{n,\ell}(x_1), \dots, \sigma_{N_\ell,\ell}(x_{N_\ell})\big)$



$$\mathrm{f}_{\text{deep}}(\boldsymbol{x}) = (\boldsymbol{\sigma}_L \circ \boldsymbol{f}_L \circ \boldsymbol{\sigma}_{L-1} \circ \cdots \circ \boldsymbol{\sigma}_2 \circ \boldsymbol{f}_2 \circ \boldsymbol{\sigma}_1 \circ \boldsymbol{f}_1)\,(\boldsymbol{x})$$

Joint learning / training ?

Constraining activation functions

- Regularization functional
  - Should not penalize simple solutions (e.g., identity or linear scaling)
  - Should impose diffentiability (for DNN to be trainable via backpropagation)
  - Should favor simplest CPWL solutions; i.e., with “sparse 2nd derivatives”

- Second total-variation of $\sigma : \mathbb{R} \to \mathbb{R}$

$$\mathrm{TV}^{(2)}(\sigma) \triangleq \|\mathrm{D}^2\sigma\|_{\mathcal{M}} = \sup_{\varphi \in \mathcal{S}(\mathbb{R}): \|\varphi\|_\infty \le 1} \langle \mathrm{D}^2\sigma, \varphi \rangle$$

- Native space for $\big(\mathcal{M}(\mathbb{R}), \mathrm{D}^2\big)$

$$\mathrm{BV}^{(2)}(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} : \|\mathrm{D}^2 f\|_{\mathcal{M}} < \infty\}$$

Representer theorem for deep neural networks

Theorem ($\mathrm{TV}^{(2)}$-optimality of deep spline networks) (Unser, JMLR 2019)

- neural network $\mathbf{f} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ with **deep structure** $(N_0, N_1, \dots, N_L)$

$$\boldsymbol{x} \mapsto \mathbf{f}(\boldsymbol{x}) = (\boldsymbol{\sigma}_L \circ \boldsymbol{\ell}_L \circ \boldsymbol{\sigma}_{L-1} \circ \cdots \circ \boldsymbol{\ell}_2 \circ \boldsymbol{\sigma}_1 \circ \boldsymbol{\ell}_1)(\boldsymbol{x})$$

- **normalized** linear transformations $\boldsymbol{\ell}_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, \boldsymbol{x} \mapsto \mathbf{U}_\ell \boldsymbol{x}$ with weights $\mathbf{U}_\ell = [\mathbf{u}_{1,\ell} \ \cdots \ \mathbf{u}_{N_\ell,\ell}]^T \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ such that $\|\mathbf{u}_{n,\ell}\| = 1$
- **free-form** activations $\boldsymbol{\sigma}_\ell = \big(\sigma_{1,\ell}, \dots, \sigma_{N_\ell,\ell}\big) : \mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$ with $\sigma_{1,\ell}, \dots, \sigma_{N_\ell,\ell} \in \mathrm{BV}^{(2)}(\mathbb{R})$

Given a series data points $(\boldsymbol{x}_m, \boldsymbol{y}_m)$ $m = 1, \dots, M$, we then define the training problem

$$\arg \min_{(\mathbf{U}_\ell), (\sigma_{n,\ell} \in \mathrm{BV}^{(2)}(\mathbb{R}))} \left(\sum_{m=1}^{M} E\big(\boldsymbol{y}_m, \mathbf{f}(\boldsymbol{x}_m)\big) + \mu \sum_{\ell=1}^{N} R_\ell(\mathbf{U}_\ell) + \lambda \sum_{\ell=1}^{L} \sum_{n=1}^{N_\ell} \mathrm{TV}^{(2)}(\sigma_{n,\ell}) \right) \quad (1)$$

- $E : \mathbb{R}^{N_L} \times \mathbb{R}^{N_L} \to \mathbb{R}^+$: arbitrary convex error function
- $R_\ell : \mathbb{R}^{N_\ell \times N_{\ell-1}} \to \mathbb{R}^+$: convex cost

If solution of (1) exists, then it is achieved by a **deep spline network** with activations of the form

$$\sigma_{n,\ell}(x) = b_{1,n,\ell} + b_{2,n,\ell} x + \sum_{k=1}^{K_{n,\ell}} a_{k,n,\ell}(x - \tau_{k,n,\ell})_+,$$

with adaptive parameters $K_{n,\ell} \le M - 2$, $\tau_{1,n,\ell}, \dots, \tau_{K_{n,\ell},n,\ell} \in \mathbb{R}$, and $b_{1,n,\ell}, b_{2,n,\ell}, a_{1,n,\ell}, \dots, a_{K_{n,\ell},n,\ell} \in \mathbb{R}$.

Outcome of representer theorem

Each neuron $\big($fixed index $(n,\ell)\big)$ is characterized by

- its number $0 \le K_{n,\ell}$ of knots (ideally, much smaller than M);
- the location $\{\tau_k = \tau_{k,n,\ell}\}_{k=1}^{K_{n,\ell}}$ of these knots (ReLU biases);
- the expansion coefficients $\mathbf{b}_{n,\ell} = (b_{1,n,\ell}, b_{2,n,\ell}) \in \mathbb{R}^2$, $\boldsymbol{a}_{n,\ell} = (a_{1,n,\ell}, \dots, a_{K,n,\ell}) \in \mathbb{R}^K$.

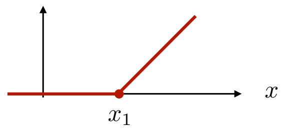
These parameters (including the number of knots) are **data-dependent** and adjusted automatically during training.

- Link with ℓ_1 minimization techniques

$$\mathrm{TV}^{(2)}\{\sigma_{n,\ell}\} = \sum_{k=1}^{K_{n,\ell}} |a_{k,n,\ell}| = \|\mathbf{a}_{n,\ell}\|_1$$

Deep spline networks: Discussion

$\mathrm{ReLU}(x; x_1) = (x - x_1)_+$



- Global optimality achieved with **spline activations**
- Justification of popular schemes / Backward compatibility
  - Standard ReLU networks $\quad (K_{n,\ell} = 1,\ \boldsymbol{b}_{n,\ell} = \mathbf{0})$ (Glorot *ICAIS* 2011) (LeCun-Bengio-Hinton *Nature* 2015)
  - Linear regression: $\lambda \to \infty \Rightarrow K_{n,\ell} = 0$
  - State-of-the-art Parametric ReLU networks $\quad (K_{n,\ell} = 1)$ (He et al. CVPR 2015)
    1 ReLU + linear term (per neuron)
  - Adaptive-piecewise linear (APL) networks $\quad (K_{n,\ell} = 5$ or $7,\ \boldsymbol{b}_{n,\ell} = \mathbf{0})$ (Agostinelli et al. 2015)

Deep spline networks (Cont'd)

- Key features
  - Direct control of complexity (number of knots): adjustment of λ
  - Ability to suppress unnecessary layers

- Challenges
  - Adaptive knots: more difficult optimization problem $\Rightarrow$ In need of novel training algorithms
  - Optimal allocation of knots
    ℓ_1-minimization with knot deletion mechanism (even for single layer)
  - Finding the tradeoff: more complex activations vs. deeper architectures

CONCLUSION: Return of the spline

- Foundations of **functional learning**
  - Functional optimization in Banach spaces (enabled by representer theorem)
  - **Hilbert spaces**: the tools of classical ML
  - **Non-convex Banach spaces**: for sparsity-promoting regularization (e.g., CS)

- Splines and machine learning
  - Traditional kernel methods are closely related to splines
    (with one knot/kernel per data point)
  - Sparse variants offer promising perspectives
  - Deep ReLU neural nets are **high-dimensional** piecewise-linear **splines**
  - Functional optimization for the streamlining of neuronal architectures
  - **Free-form** activations with TV-regularization $\Rightarrow$ ***Deep splines***

ACKNOWLEDGMENTS

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

- Dr. Julien Fageot
- Shayan Aziznejad
- 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 Jianwei Ma
- ....

erc

FunLearn

References

- 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," arXiv:2003.10112, 2020.
- Representer theorems
  - M. Unser, "A Unifying Representer Theorem for Inverse Problems and Machine Learning," *Foundations of Computational Mathematics*, vol. 21, pp. 941–960, 2021.
  - S. Aziznejad, M. Unser, "Multi-Kernel Regression with Sparsity Constraints," *SIAM Journal on Mathematics of Data Science*, vol. 3, no. 1, pp. 201-224, 2021.
- Neural networks
  - K.H. Jin, M.T. McCann, E. Froustey, M. Unser, "Deep Convolutional Neural Network for Inverse Problems in Imaging," *IEEE Trans. Image Processing*, vol. 26, no. 9, pp. 4509-4522, Sep. 2017.
  - H. Gupta, K.H. Jin, H.Q. Nguyen, M.T. McCann, M. Unser, "CNN-Based Projected Gradient Descent for Consistent CT Image Reconstruction," *IEEE Trans. Medical Imaging*, vol. 37, no. 6, pp. 1440-1453, 2018.
  - M. Unser, "A Representer Theorem for Deep Neural Networks," *J. Machine Learning Research*, vol. 20, no. 110, pp. 1-30, Jul. 2019.

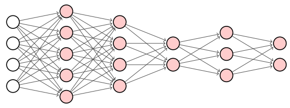
- Preprints and demos: http://bigwww.epfl.ch/

Sketch of proof

$$\min_{(\mathbf{U}_\ell),(\sigma_{n,\ell}\in \mathrm{BV}^{(2)}(\mathbb{R}))} \left(\sum_{m=1}^{M} E\big(\boldsymbol{y}_m, \mathbf{f}(\boldsymbol{x}_m)\big) + \mu \sum_{\ell=1}^{N} R_\ell(\mathbf{U}_\ell) + \lambda \sum_{\ell=1}^{L} \sum_{n=1}^{N_\ell} \mathrm{TV}^{(2)}(\sigma_{n,\ell}) \right)$$

Optimal solution $\tilde{\mathbf{f}} = \tilde{\boldsymbol{\sigma}}_L \circ \tilde{\boldsymbol{\ell}}_L \circ \tilde{\boldsymbol{\sigma}}_{L-1} \circ \cdots \circ \tilde{\boldsymbol{\ell}}_2 \circ \tilde{\boldsymbol{\sigma}}_1 \circ \tilde{\boldsymbol{\ell}}_1$ with optimized weights $\tilde{\mathbf{U}}_\ell$ and neuronal activations $\tilde{\sigma}_{n,\ell}$.

Apply "optimal" network $\tilde{\mathbf{f}}$ to each data point $\boldsymbol{x}_m$:



- Initialization (input): $\tilde{\boldsymbol{y}}_{m,0} = \boldsymbol{x}_m$.
- For $\ell = 1, \dots, L$

  $\boldsymbol{z}_{m,\ell} = (z_{1,m,\ell}, \dots, z_{N_\ell,m,\ell}) = \tilde{\mathbf{U}}_\ell \tilde{\boldsymbol{y}}_{m,\ell-1}$

  $\tilde{\boldsymbol{y}}_{m,\ell} = (\tilde{y}_{1,m,\ell}, \dots, \tilde{y}_{N_\ell,m,\ell}) \in \mathbb{R}^{N_\ell}$

  with $\tilde{y}_{n,m,\ell} = \tilde{\sigma}_{n,\ell}(z_{n,m,\ell}) \quad n = 1, \dots, N_\ell$. $\qquad \Rightarrow \quad \tilde{\mathbf{f}}(\boldsymbol{x}_m) = \tilde{\boldsymbol{y}}_{m,L}$

This fixes two terms of minimal criterion: $\sum_{m=1}^{M} E\big(\boldsymbol{y}_m, \tilde{\boldsymbol{y}}_{m,L}\big)$ and $\sum_{\ell=1}^{L} R_\ell(\tilde{\mathbf{U}}_\ell)$.

$\tilde{\mathbf{f}}$ achieves global optimum

$$\Leftrightarrow \quad \tilde{\sigma}_{n,\ell} = \arg \min_{f \in \mathrm{BV}^{(2)}(\mathbb{R})} \|\mathrm{D}^2 f\|_{\mathcal{M}} \quad \text{s.t.} \quad f(z_{n,m,\ell}) = \tilde{y}_{n,m,\ell}, \;\; m = 1, \dots, M$$