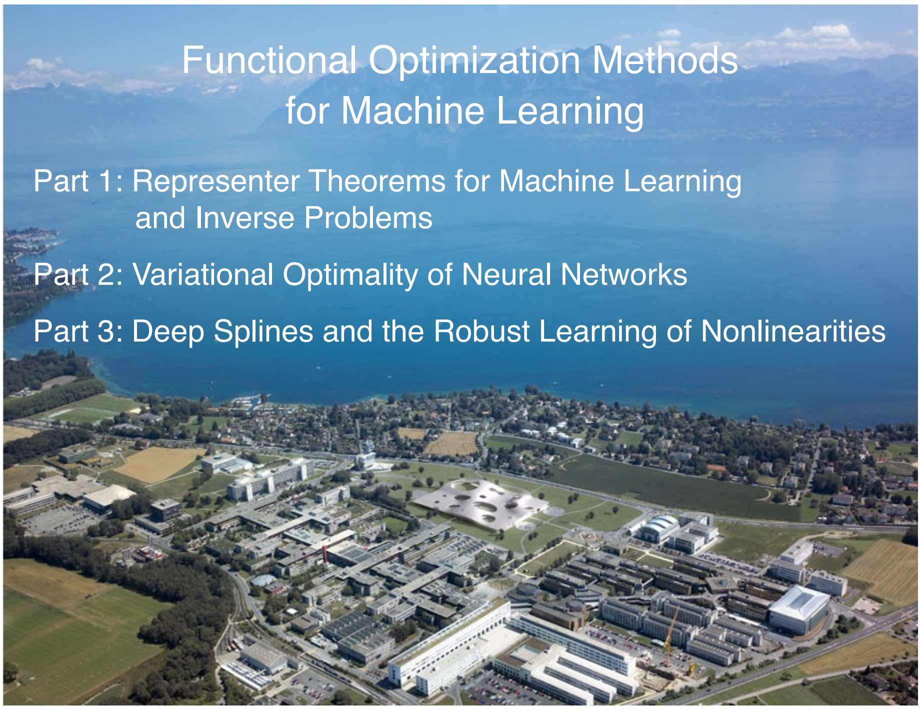
EPFL



EPFL

CIBM
Center for Biomedical Imaging

Section: Mathematical Imaging

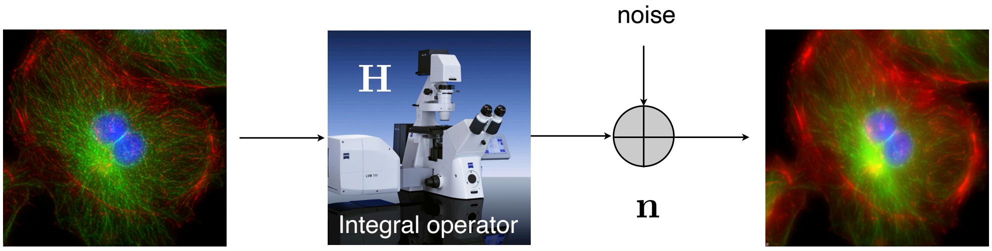
Part 1: Representer Theorems for Machine Learning and Inverse Problems

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



$\mathbf{s}$

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

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

- Regularization of ill-posed inverse problem

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

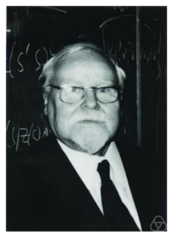
Linear inverse problems (20th century theory)

- Dealing with **ill-posed problems**: Tikhonov **regularization**

$\mathcal{R}(\mathbf{s}) = \|\mathbf{L}\mathbf{s}\|_2^2$: regularization (or smoothness) functional

$\mathbf{L}$: regularization operator (i.e., Gradient)

$$\min_{\mathbf{s}} \mathcal{R}(\mathbf{s}) \quad \text{subject to} \quad \|\mathbf{y}-\mathbf{H}\mathbf{s}\|_2^2 \le \sigma^2$$



Andrey N. Tikhonov (1906-1993)

- Equivalent variational problem

$$\mathbf{s}^\star = \arg\min \underbrace{\|\mathbf{y}-\mathbf{H}\mathbf{s}\|_2^2}_{\text{data consistency}} + \underbrace{\lambda\|\mathbf{L}\mathbf{s}\|_2^2}_{\text{regularization}}$$

Formal linear solution: $\mathbf{s} = (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{L}^T\mathbf{L})^{-1}\mathbf{H}^T\mathbf{y} = \mathbf{R}_\lambda \cdot \mathbf{y}$

Interpretation: "filtered" backprojection

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)

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

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

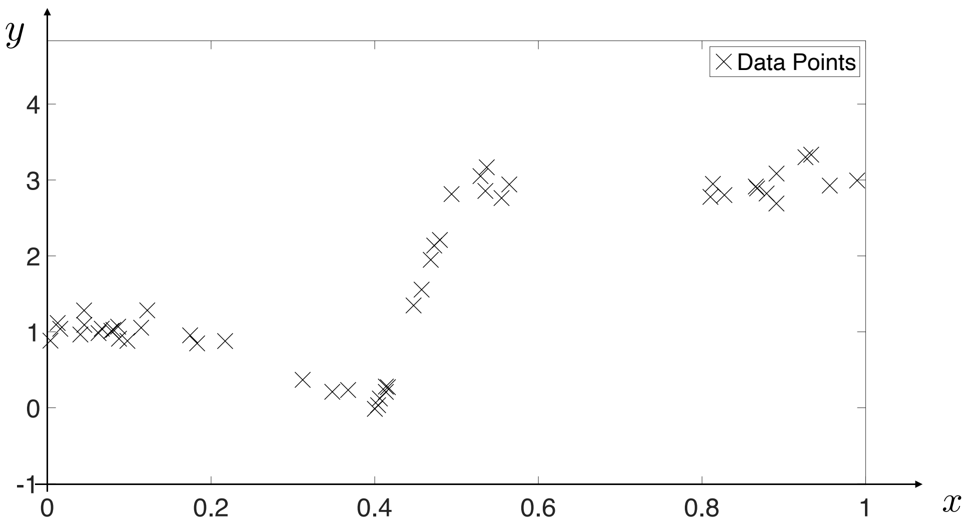
- Regularized least-squares fit (theory of RKHS)

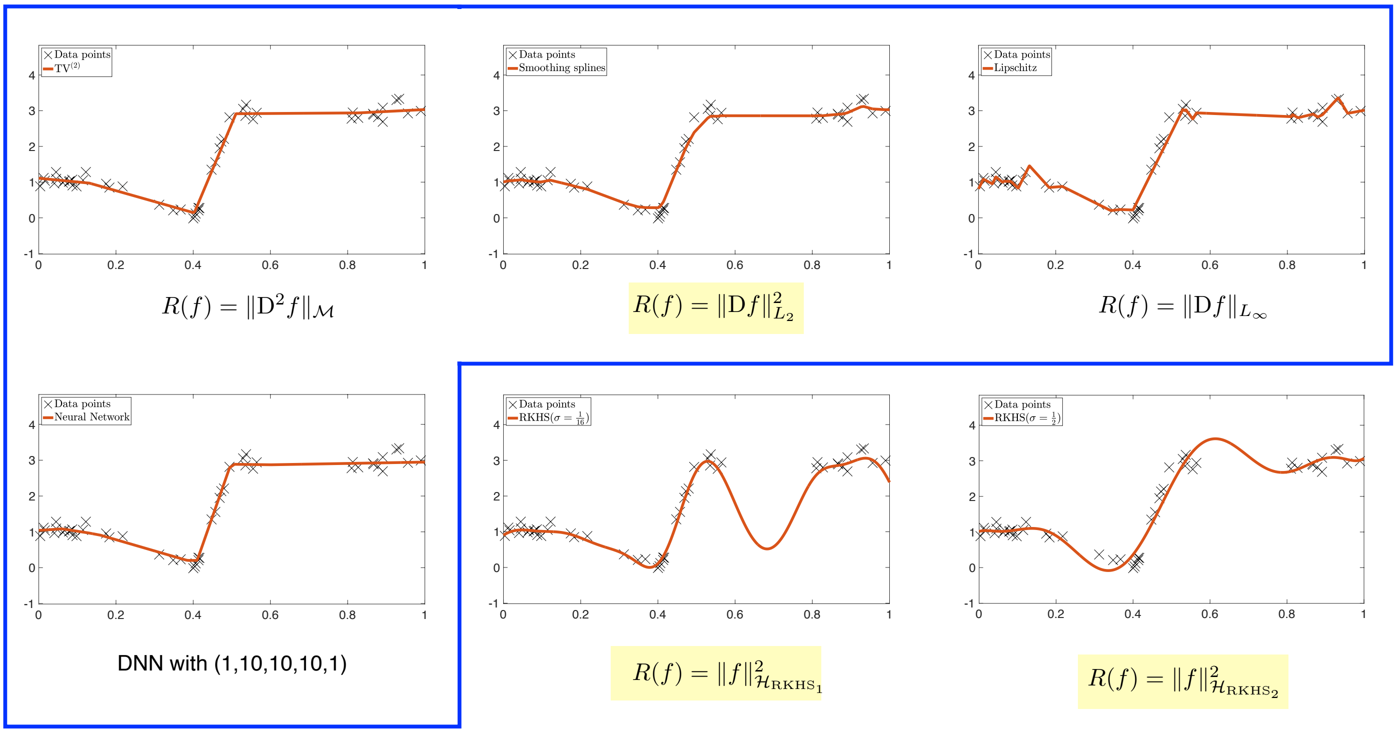
  $$f_{\mathrm{RKHS}} = \arg\min_{f \in \mathcal{H}} \left(\sum_{m=1}^{M} |y_m - f(\boldsymbol{x}_m)|^2 + \lambda R(f) \right) \quad \text{with} \quad R(f) = \|f\|^2_{\mathcal{H}}$$

  $\Rightarrow$ kernel estimator

  (Wahba 1990; Schölkopf 2001)

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

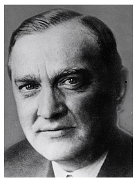




OUTLINE

- **Introduction** ✔
  - Learning as an inverse problem
- **Foundations of functional learning**
  - Banach spaces and duality mappings
  - **Unifying representer theorem** NEW
- **From classical to modern regularization-based techniques**
  - Kernel methods
  - **Sparse kernels**
  - **Sparse adaptive splines** NEW
- **Splines are universal solutions of linear inverse problems**
- **Connection with neural networks**
  - Limit behaviour of univariate shallow networks
  - **Learning free-form activations**

General notion of Banach space



Stefan Banach (1892-1945)

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

Cauchy sequence of functions $(\varphi_i)_{i\in\mathbb{N}}$ in $\mathcal{X}$:
for any $\epsilon > 0, \exists n_\epsilon$ s. t. $\|\varphi_i - \varphi_j\|_{\mathcal{X}} < \epsilon$ for all $i, j > n_\epsilon$

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 Cauchy 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**:

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

- Generic duality bound

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

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

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

Riesz conjugate for Hilbert spaces

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

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

Frigyes Riesz (1880-1956)

- Definition

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

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

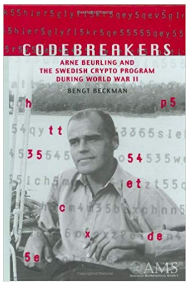
(sharp duality bound)

+

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

$(\mathcal{H}')' = \mathcal{H}$ (reflexivity)

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.

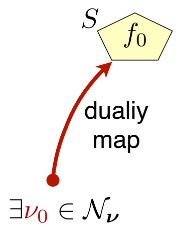
General abstract representer theorem

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 2020)

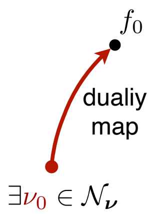
General representer theorem (Cont'd)

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 $\mathcal{X} \to \mathcal{X}'$.

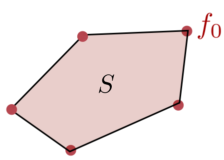
(Unser, FoCM 2020)



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

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

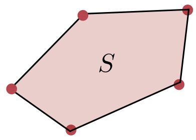
for some $K_0 \leq M$, $c_1, \dots, c_{K_0} \in \mathbb{R}$, where $e_1, \dots, e_{K_0} \in \mathcal{X}'$ are some extreme points of the unit ball $B_{\mathcal{X}'} = \{x \in \mathcal{X} : \|x\|_{\mathcal{X}'} \leq 1\}$.



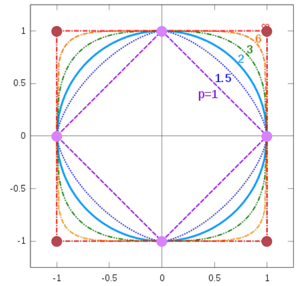
(Boyer-Chambolle-De Castro-Duval-De Gournay-Weiss, SIAM J. Optimization, 2019)

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**
  - **Lipschitz splines**

- Splines are universal solutions of linear inverse problems
- Connection with neural networks

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** $h(\cdot, \boldsymbol{x}_0) = \big(\delta(\cdot - \boldsymbol{x}_0)\big)^* \in \mathcal{H}$ when indexed by $\boldsymbol{x}_0$ is called the **reproducing kernel** of $\mathcal{H}$.

- Learning problem

Given the data $(\boldsymbol{x}_m, y_m)_{m=1}^M$ 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**

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

- 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 $\mathrm{L}^{-1}\mathrm{L}^{-1*} = (\mathrm{L}^*\mathrm{L})^{-1}$

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

- 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)

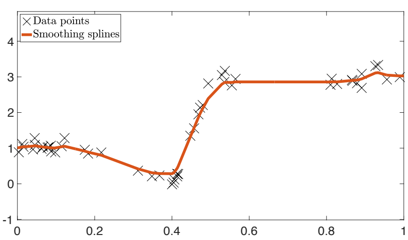
- Smoothness regularization (spline semi-norm)

$R(f) = \|\mathrm{D}f\|_{L_2}^2$ 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 $(\mathrm{D}^*\mathrm{D})^{-1}$: $\quad h(x) = \mathcal{F}^{-1}\left\{\frac{1}{|\omega|^2}\right\}(x) = \frac{1}{2}|x|$



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

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

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

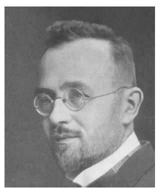
Banach isometry

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

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

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, $\|\mathrm{L}f_0\|_{\mathcal{M}} = \sum_{k=1}^{K_0} |a_k| = \|\boldsymbol{a}\|_{\ell_1}$.

RKHS vs. sparse kernel expansions (LSI)

$$\min_{f \in L_{2,\mathrm{L}}(\mathbb{R}^d)} \left(\sum_{m=1}^{M} E_m\big(y_m, f(\boldsymbol{x}_m)\big) + \lambda \|\mathrm{L}f\|_{L_2}^2 \right)$$

$$\Rightarrow \quad f_{\mathrm{RKHS}}(\boldsymbol{x}) = \sum_{m=1}^{M} a_m h_{\mathrm{PD}}(\boldsymbol{x} - \boldsymbol{x}_m)$$

Quadratic energy: $\|\mathrm{L}f_{\mathrm{RKHS}}\|_{L_2}^2 = \boldsymbol{a}^T \mathbf{G} \boldsymbol{a}$

Positive-definite kernel:

$$h_{\mathrm{PD}}(\boldsymbol{x}) = \mathcal{F}^{-1}\left\{\frac{1}{|\widehat{L}(\boldsymbol{\omega})|^2}\right\}(\boldsymbol{x})$$

Gram matrix: $[\mathbf{G}]_{m,n} = h_{\mathrm{PD}}(\boldsymbol{x}_m - \boldsymbol{x}_n)$

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

$$\Rightarrow \quad f_{\mathrm{sparse}}(\boldsymbol{x}) = \sum_{k=1}^{K_0} a_k h_{\mathrm{LSI}}(\boldsymbol{x} - \boldsymbol{\tau}_k)$$

Sparsity-promoting energy: $\|\mathrm{L}f_{\mathrm{sparse}}\|_{\mathcal{M}} = \|\boldsymbol{a}\|_{\ell_1}$

Admissible kernel:

$$h_{\mathrm{LSI}}(\boldsymbol{x}) = \mathcal{F}^{-1}\left\{\frac{1}{\widehat{L}(\boldsymbol{\omega})}\right\}(\boldsymbol{x})$$

Adaptive parameters: $K_0 \le M$, $\boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_{K_0} \in \mathbb{R}^d$

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)

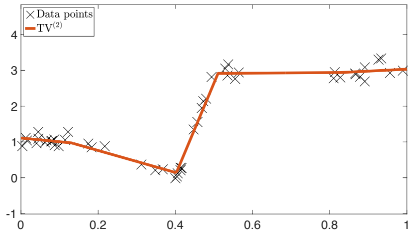
- Sparsity-promoting regularization

  $R(f) = \|\mathrm{D}^2 f\|_{\mathcal{M}}$ **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 JCAM 2022)

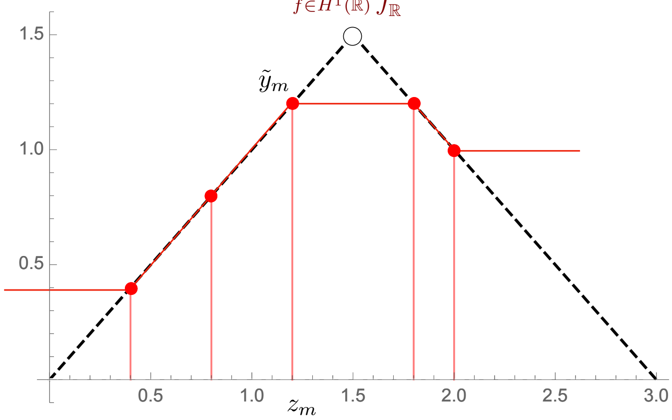
- Solution = **linear spline** with (few) **adaptive** knots at $\tau_1, \dots, \tau_{K_0}$

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

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)

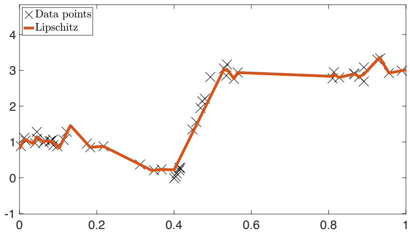
6. Lipschitz splines

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

- 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 $e_k(x) = \pm 1$



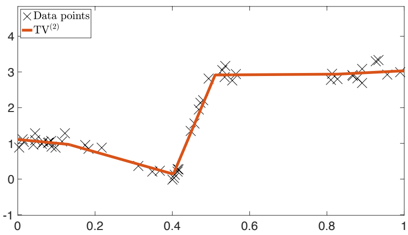
- Direct-sum Banach topology: $W^1_\infty(\mathbb{R}) = \mathcal{U}_\mathrm{D} \oplus \mathcal{N}_\mathrm{D}$

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

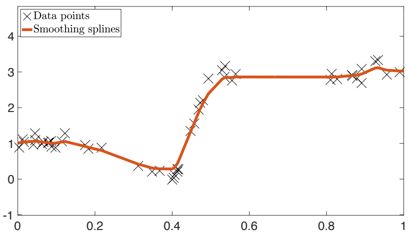
$u_k = \mathrm{D}^{-1} e_k(x) = \int_{-\infty}^{x} e_k(t)\mathrm{d}t + C_k$: linear spline with binary slope (± 1)

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

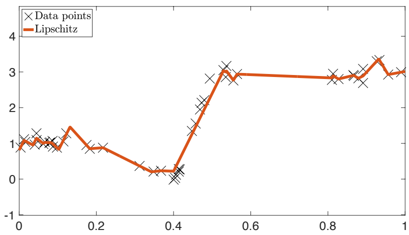
$$f_0(x) = b_0 + \sum_{k=1}^{K_0} a_k u_k(x)$$



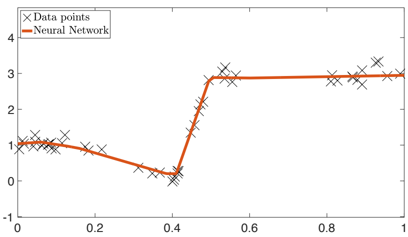
$R(f) = \|\mathrm{D}^2 f\|_\mathcal{M}$



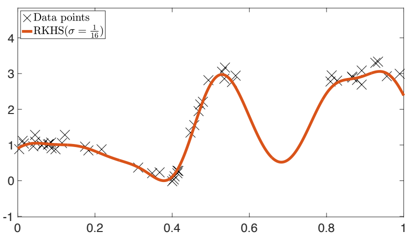
$R(f) = \|\mathrm{D}f\|^2_{L_2}$



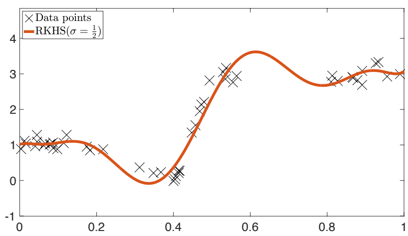
$R(f) = \|\mathrm{D}f\|_{L_\infty}$



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



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



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

OUTLINE

- Introduction ✔
- Foundations of functional learning ✔
- From classical to modern regularization-based techniques ✔
- **Splines are universal solutions of linear inverse problems**
  - Splines and operators
  - Representer theorem for gTV regularization
- Connection with neural networks

Splines and operators

Definition

A linear operator $\mathrm{L}: \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} \supseteq \mathcal{S}(\mathbb{R}^d)$ and $\mathcal{Y}$ are appropriate (dense) subspaces of $\mathcal{S}'(\mathbb{R}^d)$, is called **spline-admissible** if

1. it is linear shift-invariant (LSI);
2. its null space $\mathcal{N}_{\mathrm{L}} = \{p \in \mathcal{X} : \mathrm{L}\{p\} = 0\}$ is finite-dimensional of size N_0;
3. there exists a function $\rho_{\mathrm{L}} : \mathbb{R}^d \to \mathbb{R}$ of slow growth (Green's function of L) such that $\mathrm{L}\{\rho_{\mathrm{L}}\} = \delta$.

- Example of admissible operator

$$\mathrm{D}^n = \frac{\mathrm{d}^n}{\mathrm{d}x^n}$$

with $\rho_{\mathrm{D}^n}(x) = \frac{x_+^{n-1}}{(n-1)!}$ and $\mathcal{N}_{\mathrm{D}^n} = \mathrm{span}\left\{\frac{x^{m-1}}{(m-1)!}\right\}_{m=1}^{n}$

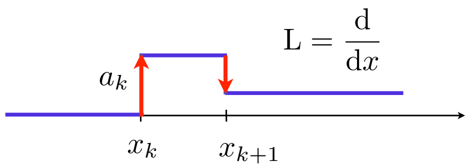
Splines are analog, but intrinsically sparse

L: spline-admissible operator; i.e., LSI and quasi-invertible

δ: Dirac distribution

Definition
The function $s : \mathbb{R}^d \to \mathbb{R}$ (possibly of slow growth) is a **nonuniform** L**-spline** with **knots** $\{\boldsymbol{x}_k\}_{k \in S}$

$$\Leftrightarrow \qquad \mathrm{L}s = \sum_{k \in S} a_k \delta(\cdot - \boldsymbol{x}_k) \ = w \ : \ \text{spline's innovation}$$



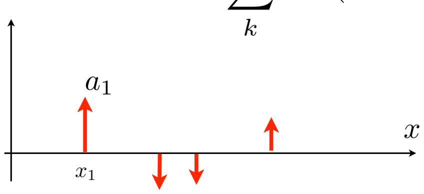
Spline theory: (Schultz-Varga, 1967)

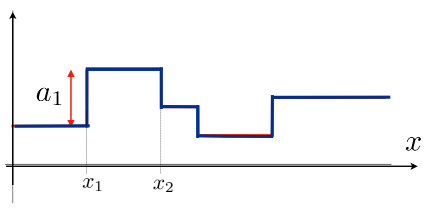
Formal spline synthesis

L: spline admissible operator (LSI)

- Finite-dimensional null space: $\mathcal{N}_{\mathrm{L}} = \mathrm{span}\{p_n\}_{n=1}^{N_0}$
- Green's function $\rho_{\mathrm{L}} : \mathbb{R}^d \to \mathbb{R}$ such that $\mathrm{L}\{\rho_{\mathrm{L}}\} = \delta$

Spline's innovation: $\quad w = \sum_k a_k \delta(\cdot - \boldsymbol{x}_k)$





$$\Rightarrow \quad s(\boldsymbol{x}) = \sum_k a_k \rho_{\mathrm{L}}(\boldsymbol{x} - \boldsymbol{x}_k) + \boxed{\sum_{n=1}^{N_0} b_n p_n(\boldsymbol{x})}$$

Requires specification of boundary conditions

Representer theorem for gTV regularization

- L: spline-admissible operator with null space $\mathcal{N}_{\mathrm{L}} = \mathrm{span}\{p_n\}_{n=1}^{N_0}$
- gTV semi-norm: $\|\mathrm{L}\{s\}\|_{\mathcal{M}} = \sup_{\|\varphi\|_\infty \le 1} \langle \mathrm{L}\{s\}, \varphi\rangle$
- Measurement functionals $h_m : \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) \to \mathbb{R}$ (weak$*$-continuous)

$$\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) = \{f \in \mathcal{S}'(\mathbb{R}^d) : \|\mathrm{L}f\|_{\mathcal{M}} < \infty\}$$

(P1) $$\arg \min_{f \in \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)} \left(\sum_{m=1}^{M} |y_m - \langle h_m, f\rangle|^2 + \lambda \|\mathrm{L}f\|_{\mathcal{M}} \right)$$

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

$\boldsymbol{\nu} : \mathcal{M}_{\mathrm{L}} \to \mathbb{R}^M$ with $\boldsymbol{\nu}(f) = \big(\langle h_1, f\rangle, \dots, \langle h_M, f\rangle\big)$

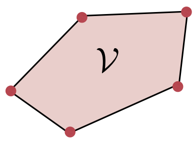
(P1') $$\arg \min_{f \in \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)} \big(E(\boldsymbol{y}, \boldsymbol{\nu}(f)) + \lambda \|\mathrm{L}f\|_{\mathcal{M}} \big)$$

Representer theorem for gTV-regularization

The extreme points of (P1') are **non-uniform** L**-spline** of the form

$$f_{\mathrm{spline}}(\boldsymbol{x}) = \sum_{k=1}^{K_{\mathrm{knots}}} a_k \rho_{\mathrm{L}}(\boldsymbol{x} - \boldsymbol{x}_k) + \sum_{n=1}^{N_0} b_n p_n(\boldsymbol{x})$$

with ρ_{L} such that $\mathrm{L}\{\rho_{\mathrm{L}}\} = \delta$, $K_{\mathrm{knots}} \le M - N_0$, and $\|\mathrm{L}f_{\mathrm{spline}}\|_{\mathcal{M}} = \|\mathbf{a}\|_{\ell_1}$.



(U.-Fageot-Ward, *SIAM Review* 2017)

Example: 1D inverse problem with TV$^{(2)}$ regularization

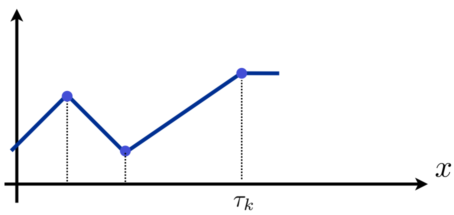
$$s_{\mathrm{spline}} = \arg \min_{s \in \mathcal{M}_{\mathrm{D}^2}(\mathbb{R})} \left(\sum_{m=1}^{M} |y_m - \langle h_m, s\rangle|^2 + \lambda \mathrm{TV}^{(2)}(s) \right)$$

- Total 2nd-variation: $\mathrm{TV}^{(2)}(s) = \sup_{\|\varphi\|_\infty \le 1} \langle \mathrm{D}^2 s, \varphi\rangle = \|\mathrm{D}^2 s\|_{\mathcal{M}}$

$\mathrm{L} = \mathrm{D}^2 = \dfrac{\mathrm{d}^2}{\mathrm{d}x^2}$ $\quad$ $\rho_{\mathrm{D}^2}(x) = (x)_+$: ReLU $\quad$ $\mathcal{N}_{\mathrm{D}^2} = \mathrm{span}\{1, x\}$

- Generic form of the solution

$$s_{\mathrm{spline}}(x) = \underbrace{b_1 + b_2 x}_{\text{no penalty}} + \sum_{k=1}^{K} a_k (x - \tau_k)_+$$



with $K < M$ and free parameters b_1, b_2 and $(a_k, \tau_k)_{k=1}^{K}$

Other spline-admissible operators

- $\mathrm{L} = \mathrm{D}^n$ (pure derivatives) (Schoenberg 1946)
  $\Rightarrow$ polynomial splines of degree $(n-1)$
- $\mathrm{L} = \mathrm{D}^n + a_{n-1}\mathrm{D}^{n-1} + \cdots + a_0\mathrm{I}$ (ordinary differential operator) (Dahmen-Micchelli 1987)
  $\Rightarrow$ exponential splines
- Fractional derivatives: $\mathrm{L} = \mathrm{D}^\gamma \quad \xleftrightarrow{\mathcal{F}} \quad (\mathrm{j}\omega)^\gamma$ (U.-Blu 2000)
  $\Rightarrow$ fractional splines
- Fractional Laplacian: $(-\Delta)^{\frac{\gamma}{2}} \quad \xleftrightarrow{\mathcal{F}} \quad \|\boldsymbol{\omega}\|^\gamma$ (Duchon 1977)
  $\Rightarrow$ polyharmonic splines
- Elliptical differential operators; e.g, $\mathrm{L} = (-\Delta + \alpha\mathrm{I})^\gamma$ (Ward-U. 2014)
  $\Rightarrow$ Sobolev splines

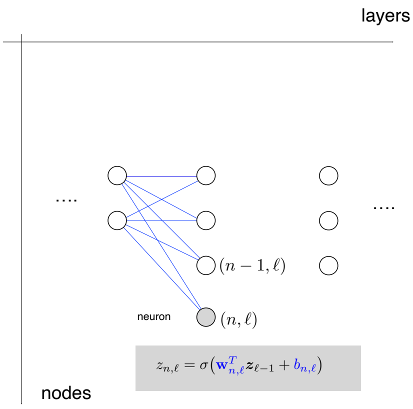
OUTLINE

- Introduction ✔
- Foundations of functional learning ✔
- From classical to modern regularization-based techniques ✔
- Splines are universal solution of inverse problems ✔
- **Connection with deep neural networks**
  - Continuous piecewise linear (CPWL) functions / splines
  - Limit behavior of shallow univariate ReLU networks
  - Representer theorem for deep neural networks

NEW

Feedforward deep neural network

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

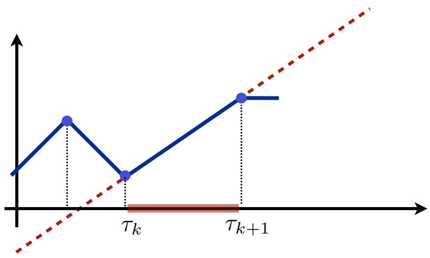


- Linear step: $\mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$
  $\boldsymbol{f}_\ell : \boldsymbol{x} \mapsto \boldsymbol{f}_\ell(\boldsymbol{x}) = \mathbf{W}_\ell \boldsymbol{x} + \mathbf{b}_\ell$
- Nonlinear step: $\mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$
  $\boldsymbol{\sigma}_\ell : \boldsymbol{x}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

Learned

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

Continuous-PieceWise Linear (CPWL) functions



- 1D: Non-uniform spline de degree 1

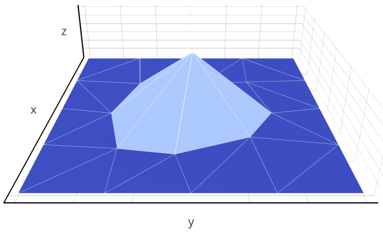
Partition: $\mathbb{R} = \bigcup_{k=0}^{K} P_k$ with $P_k = [\tau_k, \tau_{k+1})$, $\tau_0 = -\infty < \tau_1 < \cdots < \tau_K < \tau_{K+1} = +\infty$.

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

- (i) for $x \in P_k$: $f_{\text{spline}}(x) = f_k(x) \triangleq a_k x + b_k$ with $(a_k, b_k) \in \mathbb{R}^2$, $k = 0, \ldots, K$
- (ii) f_{spline} is continuous $\mathbb{R} \to \mathbb{R}$

- $f_{\text{spline}}(x) = \tilde{b}_0 + \tilde{b}_1 x + \sum_{k=1}^{K} \tilde{a}_k (x - \tau_k)_+$ $\quad$ 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.,

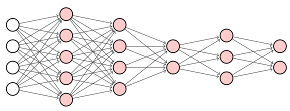
$\mathbb{R}^N = \bigcup_{k=1}^{K} P_k$ with $\mu(P_{k_1} \cap P_{k_2}) = 0$ for all $k_1 \neq k_2$

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

- (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 $\mathbf{a}_k \in \mathbb{R}^N, b_k \in \mathbb{R}$, $k = 1, \dots, K$
- (ii) f_{CPWL} is continuous $\mathbb{R}^N \to \mathbb{R}$

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

Deep ReLU neural networks are splines



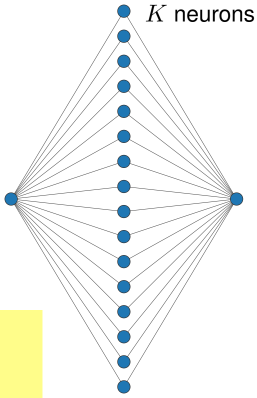
$$\mathbf{f}_{\mathrm{deep}}(\boldsymbol{x}) = (\boldsymbol{\sigma}_L \circ \boldsymbol{f}_L \circ \boldsymbol{\sigma}_{L-1} \circ \cdots \circ \boldsymbol{\sigma}_2 \circ \boldsymbol{f}_2 \circ \boldsymbol{\sigma}_1 \circ \boldsymbol{f}_1)(\boldsymbol{x})$$

- Enabling property

Composition $\mathbf{f}_2 \circ \mathbf{f}_1$ of two CPWL functions with compatible domain and range is CPWL.

- Each linear layer $\boldsymbol{f}_\ell(\boldsymbol{x}) = \mathbf{W}_\ell \boldsymbol{x} + \mathbf{b}_\ell$ is (trivially) CPWL (Montufar *NIPS* 2014)
- Each scalar neuron activation, $\sigma_{n,\ell}(x) = \mathrm{ReLU}(x)$, is CPWL (Strang *SIAM News 2018*)
  $\Rightarrow \boldsymbol{\sigma}_\ell = (\sigma_{1,\ell}, \dots, \sigma_{N_\ell,\ell})$ (pointwise nonlinearity) is CPWL
- The whole feedforward network $\mathbf{f}_{\mathrm{deep}} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ is CPWL
- 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**:

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

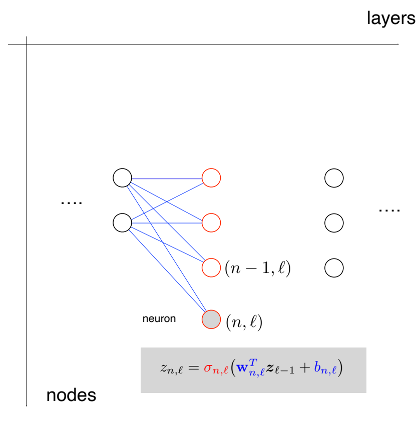
Arguments for the proof:

(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 $a_k = v_k |w_k|$.

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, ℓ), $\quad n = 1, \cdots, N_\ell$
- 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: $\mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$

  $\boldsymbol{\sigma}_\ell : \boldsymbol{x} \mapsto \boldsymbol{\sigma}_\ell(\boldsymbol{x}) = \left(\sigma_{n,\ell}(x_1), \dots, \sigma_{N_\ell,\ell}(x_{N_\ell})\right)$

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

Joint learning / training ?

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 $\boldsymbol{\sigma}_\ell=(\sigma_{1,\ell},\dots,\sigma_{N_\ell,\ell}):\mathbb{R}^{N_\ell}\to\mathbb{R}^{N_\ell}$ 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

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

- $E:\mathbb{R}^{N_L}\times\mathbb{R}^{N_L}\to\mathbb{R}^+$: 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 $b_{1,n,\ell},b_{2,n,\ell},a_{1,n,\ell},\dots,a_{K_{n,\ell},n,\ell}\in\mathbb{R}$.

CONCLUSION: Central role of splines

- 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**
  - Approximation properties of **shallow networks** are fully explained by spline theory
  - **Free-form** activations with TV-regularization ⇒ ***Deep splines***

ACKNOWLEDGMENTS

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

- Dr. Julien Fageot
- Prof. John Paul Ward
- Dr. Shayan Aziznejad
- 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 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," *Journal of Computational and Applied Mathematics*, vol. 406, paper no. 114044, 30 p., May 1, 2022.
- Representer theorems
  - M. Unser, "A Unifying Representer Theorem for Inverse Problems and Machine Learning," *Foundations of Computational Mathematics*, vol. 21, no. 4, pp. 941–960, August, 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
  - M. Unser, "A Representer Theorem for Deep Neural Networks," *J. Machine Learning Research*, vol. 20, no. 110, pp. 1-30, Jul. 2019.
  - H. P. Savarese, I. Evron, D. Soudry, and N. Srebro, "How do infinite width bounded norm networks look in function space?" in *Proc. Conf. Learn. Theory*, 2019, pp. 2667–2690.
  - R. Parhi and R. D. Nowak. The role of neural network activation functions. *IEEE Signal Processing Letters*, vol. 27, pp. 1779–1783, 2020.

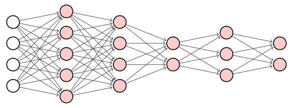
- Preprints and demos: http://bigwww.epfl.ch/

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

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