EPFL

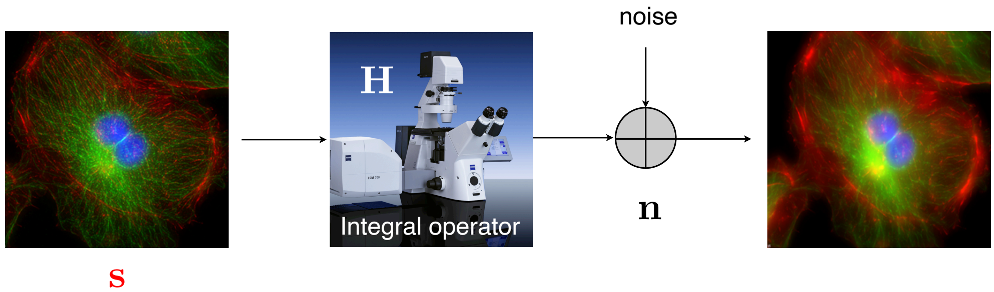
Representer theorems for machine learning and inverse problems

Michael Unser
Biomedical Imaging Group
EPFL, Lausanne, Switzerland

One World Seminar “Mathematical Methods for Arbitrary Data Sources”, June 8, 2020

Variational formulation of inverse problems

- Linear forward model



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

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

- Reconstruction as an optimization problem

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

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

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

- Regularized least-squares fit (theory of RKHS)

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

  $\Rightarrow$ kernel estimator

  (Wahba 1990; Schölkopf 2001)

RKHS representer theorem for machine learning

(P2) $\arg\min_{f \in \mathcal{H}} \left(\sum_{m=1}^{M} |y_m - f(\boldsymbol{x}_m)|^2 + \lambda \|f\|^2_{\mathcal{H}} \right)$ (Poggio-Girosi 1990)

$r_{\mathcal{H}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the (unique) **reproducing kernel** for the RKHS $\mathcal{H}$ if

- $r_{\mathcal{H}}(\cdot, \boldsymbol{x}_0) \in \mathcal{H}$ for all $\boldsymbol{x}_0 \in \mathbb{R}^d$
- $f(\boldsymbol{x}_0) = \langle r_{\mathcal{H}}(\cdot, \boldsymbol{x}_0), f \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $\boldsymbol{x}_0 \in \mathbb{R}^d$

(Aronszajn, 1950)

Formal characterization: $r_{\mathcal{H}}(\cdot, \boldsymbol{x}_0) = \mathrm{R}\{\delta(\cdot - \boldsymbol{x}_0)\} = \left(\delta(\cdot - \boldsymbol{x}_0)\right)^*$ (Riesz conjugate)

Representer theorem for L_2-regularization

The solution of (P2) has the generic parametric form: $f(\boldsymbol{x}) = \sum_{m=1}^{M} a_m r_{\mathcal{H}}(\boldsymbol{x}, \boldsymbol{x}_m)$

(de Boor 1966; Kimeldorf-Wahba 1971; Poggio-Girosi 1990)

RKHS representer theorem for machine learning

(P2') $$\arg\min_{f\in\mathcal{H}}\left(\sum_{m=1}^{M}E\big(y_m,f(\boldsymbol{x}_m)\big)+\lambda\|f\|_{\mathcal{H}}^2\right)$$ with $E:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ convex

$r_{\mathcal{H}}:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ is the (unique) **reproducing kernel** for the RKHS $\mathcal{H}$ if

- $r_{\mathcal{H}}(\cdot,\boldsymbol{x}_0)\in\mathcal{H}$ for all $\boldsymbol{x}_0\in\mathbb{R}^d$
- $f(\boldsymbol{x}_0)=\langle r_{\mathcal{H}}(\cdot,\boldsymbol{x}_0),f\rangle_{\mathcal{H}}$ for all $f\in\mathcal{H}$ and $\boldsymbol{x}_0\in\mathbb{R}^d$

(Aronszajn, 1950)

Formal characterization: $r_{\mathcal{H}}(\cdot,\boldsymbol{x}_0)=\mathrm{R}\{\delta(\cdot-\boldsymbol{x}_0)\}=\big(\delta(\cdot-\boldsymbol{x}_0)\big)^*$ (Riesz conjugate)

Representer theorem for L_2-regularization

The solution of (P2') has the generic parametric form: $$f(\boldsymbol{x})=\sum_{m=1}^{M}a_m r_{\mathcal{H}}(\boldsymbol{x},\boldsymbol{x}_m)$$

(de Boor 1966; Kimeldorf-Wahba 1971; Poggio-Girosi 1990; Schölkopf 2001)

Supports the theory of SVM, **kernel methods**, etc.

Is there a mother of all representer theorems ?

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

Classical representer theorem in machine learning:

- $\mathcal{X}'=\mathcal{H}$ is a reproducing kernel Hilbert space.
- $\boldsymbol{\nu}:\mathcal{H}'\to\mathbb{R}^M: f\mapsto\big(f(\boldsymbol{x}_1),\dots,\langle f(\boldsymbol{x}_M)\big)$ is the sampling operator.

(de Boor 1966; Kimeldorf-Wahba 1971; Poggio-Girosi 1990; Schölkopf 2001)

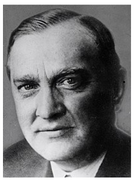
Most general set-up:

- $\mathcal{X}$ is a Banach space.
- $\boldsymbol{\nu}:\mathcal{X}\to\mathbb{R}^M: f\mapsto\big(\langle\nu_1,f\rangle,\dots,\langle\nu_M,f\rangle\big)$ is a general linear measurement operator.
- $E:\mathbb{R}^M\times\mathbb{R}^M\to\mathbb{R}^+\cup\{+\infty\}$ is a proper l.s.c. convex loss functional.
- $\psi:\mathbb{R}^+\to\mathbb{R}^+$ is some arbitrary strictly-increasing convex function.

CONTENT

- **Regularization in science and engineering** ✔
  - Inverse problems and learning
  - Representer theorem for RKHS
- **Unifying representer theorem**
  - Banach spaces and their duals
  - Duality mapping
  - Mother of all representer theorems
- **Applications**: Optimization in specific Banach spaces
  - Learning in RKHS
  - Tikhonov regularization
  - l_p-norm regularization
  - Sparse kernel expansions
  - Deep neural networks

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 $\mathcal{X}$:

$\lim_i \varphi_i = \varphi$; 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 operator $\mathrm{U} : \mathcal{X} \to \mathcal{Y}$

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**:

$$\|g\|_{\mathcal{X}'} = \sup_{f\in\mathcal{X}\setminus\{0\}} \left(\frac{\langle g, f\rangle}{\|f\|_{\mathcal{X}}}\right) = \sup_{f\in\mathcal{X}:\ \|f\|_{\mathcal{X}}\le 1} |\langle g, f\rangle|$$

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

Frigyes Riesz (1880-1956)

- Definition

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

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

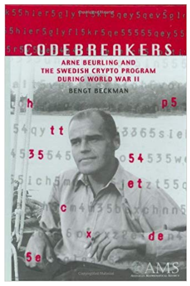
(sharp duality bound)

- Properties

+

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

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. $J(\lambda f) = \lambda J(f)$ for any $\lambda \in \mathbb{R}^+$ (homogeneity).
3. For every $f \in \mathcal{X}$, the set $J(f)$ is convex and weak-$*$ closed in $\mathcal{X}'$.
4. The duality mapping is **single-valued** if $\mathcal{X}'$ is **strictly convex**; the latter condition is also necessary if $\mathcal{X}$ is reflexive.
5. When $\mathcal{X}$ is **reflexive**, then the duality map is **bijective** if and only if both $\mathcal{X}$ and $\mathcal{X}'$ are **strictly convex**.

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

$\mathcal{X}$ 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 $\lambda \in (0,1)$.

Mother of all representer theorems

Lausanne, Christmas 2018

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

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.

Mother of all representer theorems (Cont'd)

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.

Any solution $f_0\in S\subset\mathcal{X}'$ is a $(\mathcal{X}',\mathcal{X})$-**conjugate of a common**

$$\nu_0=\sum_{m=1}^{M}a_m\nu_m\in\mathcal{N}_{\boldsymbol{\nu}}\subset\mathcal{X}$$

with suitable weights $\boldsymbol{a}\in\mathbb{R}^M$; i.e., $S\subseteq J(\nu_0)$.

If the Banach space $\mathcal{X}$ is **reflexive and strictly convex**, then the solution is **unique** with $f_0=\mathrm{J}_{\mathcal{X}}\{\nu_0\}\in\mathcal{X}'$ (Banach conjugate of ν_0). If $\mathcal{X}$ is a Hilbert space, then $f_0=\sum_{m=1}^{M}a_m\nu_m^*$ where ν_m^* is the Riesz conjugate of ν_m.

(Unser, ArXiv 2019)

CONTENT

- Regularization in science and engineering ✔
- Unifying representer theorem ✔
- **Applications**: Optimization in specific Banach spaces
  - Learning in RKHS
  - Tikhonov regularization
  - l_p-norm regularization
  - Sparsity promoting regularization
  - Sparse kernel expansions
  - Deep 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 $\delta(\cdot - \boldsymbol{x}) \in \mathcal{H}'$ for any $\boldsymbol{x} \in \mathbb{R}^d$. The corresponding unique **Hilbert conjugate** $h(\cdot, \boldsymbol{x}) = \big(\delta(\cdot - \boldsymbol{x})\big)^* \in \mathcal{H}$ when indexed by $\boldsymbol{x}$ 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 generalized representer theorem
  - $\mathcal{X} = \mathcal{H}', \ \mathcal{X}' = \mathcal{H}''$ (Reflexive Banach space)
  - $\nu_m = \delta(\cdot - \boldsymbol{x}_m)$ (Dirac sampling functionals)
  - Additive loss: $E(\boldsymbol{y}, \boldsymbol{z}) = \sum_{m=1}^{M} E_m(y_m, z_m)$

- Key observation

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

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

- Implied form of unique solution = linear kernel expansion

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

(Schölkopf representer theorem, 2001)

2. Tikhonov regularization

$\mathcal{H}$: **Hilbert space** on $\mathbb{R}^d$ with **Riesz map** $\mathrm{J}_{\mathcal{H}'} = \mathrm{R} : \mathcal{H}' \to \mathcal{H}$

- Ill-posed linear inverse problem

Measurement functionals: $\nu_1, \cdots, \nu_M \in \mathcal{H}'$

Goal: recover a function $f : \mathbb{R}^d \to \mathbb{R}$ from noisy measurements $y_m = \langle \nu_m, f \rangle + \epsilon_m$

- Formulation of reconstruction problem (penalized least-squares)

Given the data $\boldsymbol{y} \in \mathbb{R}^M$, find the function $f_0 : \mathbb{R}^d \to \mathbb{R}$ s.t.

$$f_0 = \arg\min_{f \in \mathcal{H}} \left(\sum_{m=1}^{M} |y_m - \langle \nu_m, f \rangle|^2 + \lambda \|f\|_{\mathcal{H}}^2 \right)$$

$\lambda \in \mathbb{R}^+$: adjustable regularization parameter.

Tikhonov regularization: closed-form solution

- Application of generalized representer theorem
  - $\mathcal{X} = \mathcal{H}', \ \mathcal{X}' = \mathcal{H}'' = \mathcal{H}$ (Hilbert space)
  - Measurement functionals: $\nu_m \in \mathcal{H}', m = 1, \dots, M$
  - Conjugate functions: $\varphi_m = \nu_m^* = \mathrm{R}\{\nu_m\} \in \mathcal{H}$
  - $\psi(t) = \lambda |t|^2$ (convex)

$$\Rightarrow \quad f_0 = \sum_{m=1}^{M} a_m \varphi_m \quad \in \mathrm{span}\{\varphi_m\}$$

- Optimal discretization: "*the miraculous simplification*"
  - System matrix $\mathbf{H} \in \mathbb{R}^{M \times M}$ = **Gram matrix** (symmetric, positive-definite)

    $$[\mathbf{H}]_{m,n} = \langle \nu_m, \varphi_n \rangle = \langle \nu_m, \nu_n^* \rangle = \langle \nu_m^*, \nu_n^* \rangle_{\mathcal{H}} = \langle \varphi_m, \varphi_n \rangle_{\mathcal{H}}$$
  - $f = \sum_{m=1}^{M} a_m \varphi_m \quad \Rightarrow \quad \boldsymbol{\nu}(f) = \mathbf{H}\boldsymbol{a}, \quad \|f\|_{\mathcal{H}}^2 = \boldsymbol{a}^T \mathbf{H} \boldsymbol{a}$

$$\Rightarrow \quad \boldsymbol{a}_{\mathrm{opt}} = \arg \min_{\boldsymbol{a} \in \mathbb{R}^M} \left(\|\boldsymbol{y} - \mathbf{H}\boldsymbol{a}\|_2^2 + \lambda \|\mathbf{H}\boldsymbol{a}\|_2^2 \right) = (\mathbf{H} + \lambda \mathbf{I})^{-1} \boldsymbol{y}$$

3. ℓ_p-norm regularization

- Finite-dimensional setup (CS)
  - Goal: Recover $\boldsymbol{s} = (s_n) \in \mathbb{R}^N$ from a set of corrupted linear measurements
    $y_m = \boldsymbol{h}_m^T \boldsymbol{s} + \epsilon_m, \ m = 1, \dots, M$
  - **Compressed sensing** scenario: $M \ll N$
  - Strategy: Try to favor **sparse** solutions
- Formulation of reconstruction task
  - Data $\boldsymbol{y} \in \mathbb{R}^M$
  - System matrix $\mathbf{H} = [\boldsymbol{h}_1 \ \boldsymbol{h}_2 \ \cdots \ \mathbf{h}_M]^T \in \mathbb{R}^{M \times N}$
  - Minimization problem with $p > 0$ small

$$\boldsymbol{s} = \arg \min_{\boldsymbol{x} \in \mathbb{R}^N} \left(E(\boldsymbol{y}, \mathbf{H}\boldsymbol{x}) + \lambda \|\boldsymbol{x}\|_{\ell_p}^p \right)$$

$\lambda \in \mathbb{R}^+$: adjustable regularization parameter

ℓ_p-norm regularization (Cont'd)

- Application of general representer theorem
  - $\mathcal{X} = (\mathbb{R}^N, \|\cdot\|_{\ell_q})$, $\mathcal{X}' = (\mathbb{R}^N, \|\cdot\|_{\ell_p})$ with $\frac{1}{p} + \frac{1}{q} = 1$
  - Hölder inequality: $|\langle \boldsymbol{u}, \boldsymbol{v}\rangle| \le \|\boldsymbol{u}\|_{\ell_p} \|\boldsymbol{v}\|_{\ell_q}$
  - $\psi(x) = \lambda |x|^p$ is convex for $p \ge 1$
  - $\|\cdot\|_{\ell_p}$ and $\|\cdot\|_{\ell_q}$ norms are strictly convex for $p \in (1, \infty)$ $\Rightarrow$ **unique solution**
  - Known q-to-p duality map: $[\boldsymbol{v}^*]_n = \frac{|v_n|^{q-1}}{\|\boldsymbol{v}\|_{\ell_q}^{q-2}} \mathrm{sign}(v_n)$

- Parametric form of the solution:

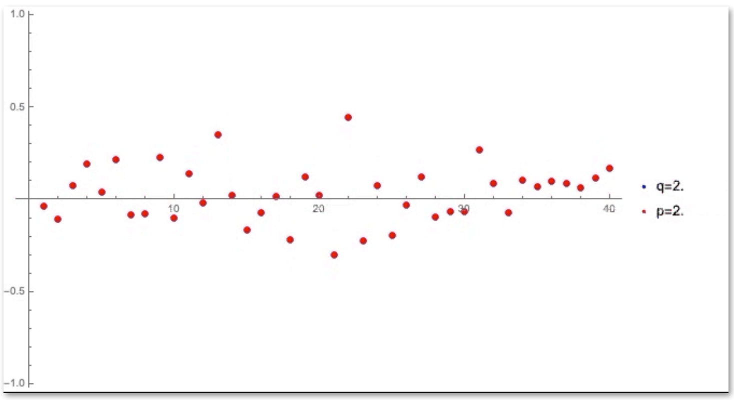
$$[\boldsymbol{s}]_n = \frac{\left|[\mathbf{H}^T \boldsymbol{a}]_n)\right|^{q-1}}{\|\mathbf{H}^T \boldsymbol{a}\|_{\ell_q}^{q-2}} \mathrm{sign}\big([\mathbf{H}^T \boldsymbol{a}]_n\big)$$

with parameter vector $\boldsymbol{a} \in \mathbb{R}^M$

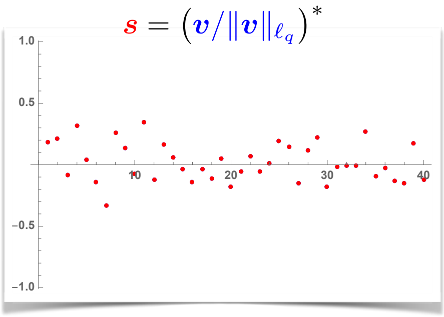
Qualitative effect of Banach conjugation

$$\mathrm{J}_{\ell_q(\mathbb{Z})} : \ell_q(\mathbb{Z}) \to \ell_p(\mathbb{Z}) \qquad v_n^* = \frac{|v_n|^{q-1}}{\|\boldsymbol{v}\|_{\ell_q}^{q-2}} \mathrm{sign}(v_n)$$

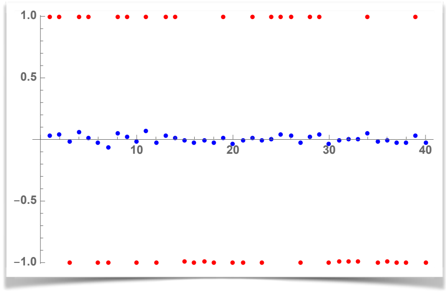
$$\boldsymbol{s} = \big(\boldsymbol{v}/\|\boldsymbol{v}\|_{\ell_q}\big)^*$$



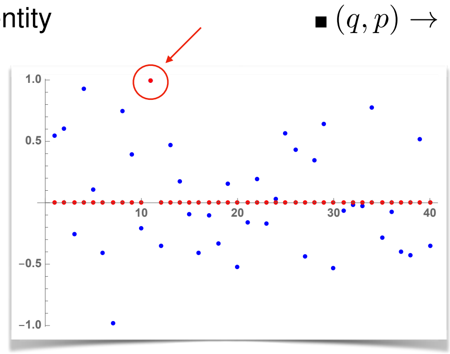
Qualitative effect of Banach conjugation



■ $(q,p) = (2,2)$: identity



■ $(q,p) \to (1,\infty)$: saturation of v^*

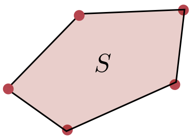


■ $(q,p) \to (\infty,1)$: sparsification of v^*

4. Sparsity promoting regularization

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

- Cases where the solution set is not necessarily unique
  - $\mathcal{X}'$ is non-reflexive, non-strictly convex; e.g., $\mathcal{X}' = \ell_1(\mathbb{Z})$
  - Representer theorem $\Rightarrow$ S is convex, weak* **compact**
  - Krein-Milman theorem: S is the convex hull of its **extreme points**



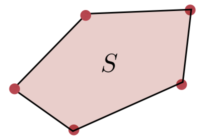
Theorem
All extreme points f_0 of S can be expressed as

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

for some $1 \leq K_0 \leq M$ where the e_k are some **extreme points** of the unit “regularization” ball $B_{\mathcal{X}'} = \{f \in \mathcal{X}' : \|f\|_{\mathcal{X}'} \leq 1\}$ and $\boldsymbol{a} = (a_1, \cdots, a_{K_0}) \in \mathbb{R}^{K_0}$.

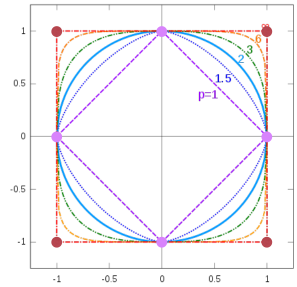
(Boyer-Chambolle-De Castro-Duval-De Gournay-Weiss, arXiv:1806.09810, 2019)

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 $\ell_p(\mathbb{Z})$

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

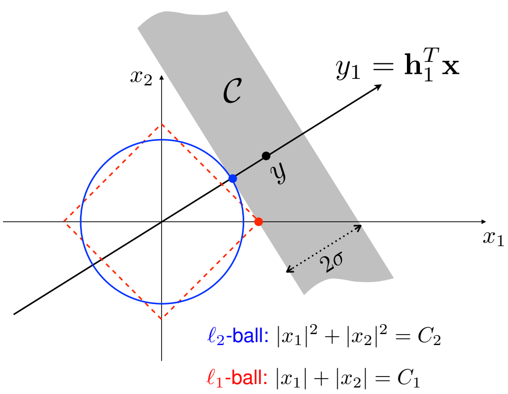
Definition of ***strictly convex*ity**: all boundary points are extreme !!!

Geometry of l_2 vs. l_1 minimization

- Prototypical inverse problem

$$\min_{\mathbf{x}} \left\{ \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_{\ell_2}^2 + \lambda \|\mathbf{x}\|_{\ell_2}^2 \right\} \Leftrightarrow \min_{\mathbf{x}} \|\mathbf{x}\|_{\ell_2} \text{ subject to } \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_{\ell_2}^2 \le \sigma^2$$

$$\min_{\mathbf{x}} \left\{ \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_{\ell_2}^2 + \lambda \|\mathbf{x}\|_{\ell_1} \right\} \Leftrightarrow \min_{\mathbf{x}} \|\mathbf{x}\|_{\ell_1} \text{ subject to } \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_{\ell_2}^2 \le \sigma^2$$

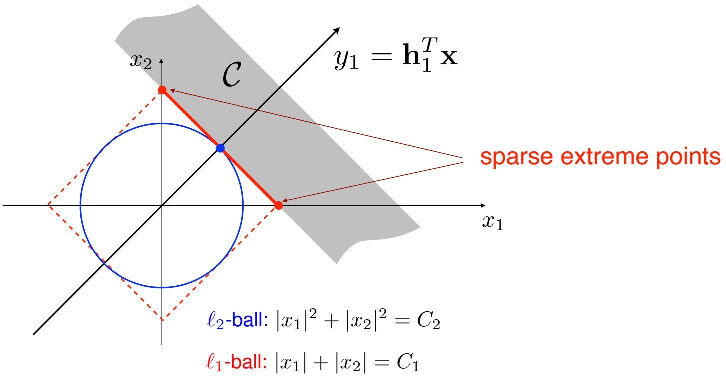


Geometry of l_2 vs. l_1 minimization

- Prototypical inverse problem

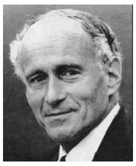
$$\min_{\mathbf{x}} \left\{ \|\mathbf{y}-\mathbf{H}\mathbf{x}\|_{\ell_2}^2 + \lambda \|\mathbf{x}\|_{\ell_2}^2 \right\} \quad \Leftrightarrow \quad \min_{\mathbf{x}} \|\mathbf{x}\|_{\ell_2} \text{ subject to } \|\mathbf{y}-\mathbf{H}\mathbf{x}\|_{\ell_2}^2 \le \sigma^2$$

$$\min_{\mathbf{x}} \left\{ \|\mathbf{y}-\mathbf{H}\mathbf{x}\|_{\ell_2}^2 + \lambda \|\mathbf{x}\|_{\ell_1} \right\} \quad \Leftrightarrow \quad \min_{\mathbf{x}} \|\mathbf{x}\|_{\ell_1} \text{ subject to } \|\mathbf{y}-\mathbf{H}\mathbf{x}\|_{\ell_2}^2 \le \sigma^2$$



Configuration for **non-unique** ℓ_1 solution

5. Sparse kernel expansions



Laurent Schwartz (1915-2002)

- Context
  - $\mathcal{S}(\mathbb{R}^d)$: Schwartz's space of smooth and rapidly decaying functions on $\mathbb{R}^d$
  - $\mathcal{S}'(\mathbb{R}^d)$: the space of tempered distributions
  - Regularization operator $\mathrm{L} : \mathcal{S}'(\mathbb{R}^d) \xrightarrow{\text{c.}} \mathcal{S}'(\mathbb{R}^d)$
  - Inverse operator $\mathrm{L}^{-1} : \mathcal{S}'(\mathbb{R}^d) \xrightarrow{\text{c.}} \mathcal{S}'(\mathbb{R}^d)$
  - Bivariate kernel: $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

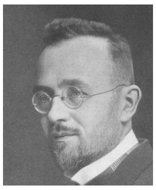
$$\mathrm{L}^{-1}\{\varphi\} = \int_{\mathbb{R}^d} h(\cdot, \boldsymbol{y})\varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$$

Schwartz kernel

- Native Banach space for $\left(\mathrm{L}, \mathcal{M}(\mathbb{R}^d)\right)$

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

Isometry with space of Radon measures



Johann Radon (1887-1956)

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

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

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

- Basic isometries

$$\mathrm{L} : \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d) \to \mathcal{M}(\mathbb{R}^d)$$
$$\mathrm{L}^{-1} : \mathcal{M}(\mathbb{R}^d) \to \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$$

$$\mathrm{L}^{-1} : \varphi \mapsto \int_{\mathbb{R}^d} h(\cdot, \boldsymbol{y})\varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$$

- **Extreme points** of unit ball in $\mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)$:

$$u_k = \mathrm{L}^{-1}\{e_k\} = \pm\mathrm{L}^{-1}\{\delta(\cdot - \boldsymbol{\tau}_k)\} = \pm h(\cdot, \boldsymbol{\tau}_k)$$

Sparse kernel expansions (Cont'd)

$$\mathrm{L}^{-1} : \varphi \mapsto \int_{\mathbb{R}^d} h(\cdot, \boldsymbol{y})\varphi(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$$

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

Theorem
All extreme points f_0 of S can be expressed as

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

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

Special case: Translation-invariant kernels

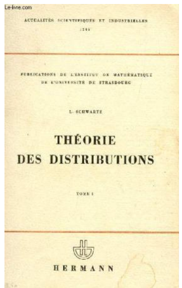
- Linear-shift invariant (LSI) setting
  - LSI operator L with **frequency response** $\widehat{L}(\boldsymbol{\omega}) = \mathcal{F}\{\mathrm{L}\{\delta\}\}(\boldsymbol{\omega})$
  - LSI inverse operator $\mathrm{L}^{-1}: \varphi \mapsto h_{\mathrm{LSI}} * \varphi$
  - Translation-invariant kernel: $h(\boldsymbol{x}, \boldsymbol{\tau}) = h_{\mathrm{LSI}}(\boldsymbol{x} - \boldsymbol{\tau})$
- Determination of the kernel: $h_{\mathrm{LSI}}(\boldsymbol{x}) = \mathcal{F}^{-1}\left\{\frac{1}{\widehat{L}(\boldsymbol{\omega})}\right\}(\boldsymbol{x})$
- Determination of the regularization operator

$$\widehat{L}(\boldsymbol{\omega}) = \frac{1}{\widehat{h}_{\mathrm{LSI}}(\boldsymbol{\omega})}$$

$\mathrm{L}: \mathcal{S}'(\mathbb{R}^d) \xrightarrow{\text{c.}} \mathcal{S}'(\mathbb{R}^d) \quad \Leftrightarrow \quad \widehat{L}(\boldsymbol{\omega})$ smooth and slowly growing

Example of admissible kernels:

$h_{\mathrm{LSI}}(\boldsymbol{x}) = \exp\left(-\|\boldsymbol{x}\|^{\alpha}\right)$ with $\alpha \in (0, 2)$



RKHS vs. sparse kernel expansions (LSI)

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

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

Positive-definite kernel:

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

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

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

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

Admissible kernel:

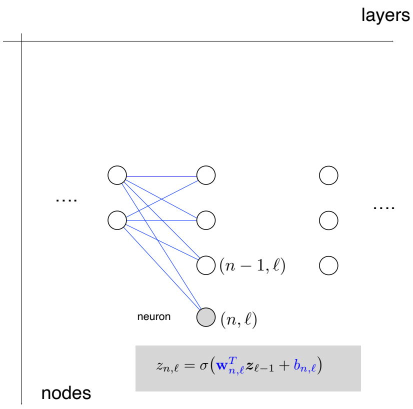
$$h_{\mathrm{LSI}}(\boldsymbol{x}) = \mathcal{F}^{-1}\left\{\frac{1}{\widehat{L}(\boldsymbol{\omega})}\right\}(\boldsymbol{x})$$

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

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

6. 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)



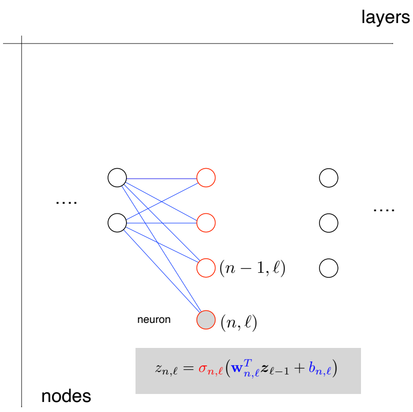
- Linear step: $\mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$
  $\boldsymbol{f}_\ell : \boldsymbol{x} \mapsto \boldsymbol{f}_\ell(\boldsymbol{x}) = \mathbf{W}_\ell \boldsymbol{x} + \mathbf{b}_\ell$
- Nonlinear step: $\mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$
  $\boldsymbol{\sigma}_\ell : \boldsymbol{x}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

Learned

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

Refinement: free-form activation functions

- Layers: $\ell = 1, \dots, L$
- Deep structure descriptor: $(N_0, N_1, \cdots, N_L)$
- Neuron or node index: $(n, \ell), \quad n = 1, \cdots, N_\ell$
- 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: $\mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$
  $\boldsymbol{\sigma}_\ell : \boldsymbol{x} \mapsto \boldsymbol{\sigma}_\ell(\boldsymbol{x}) = \big(\sigma_{n,\ell}(x_1), \dots, \sigma_{N_\ell,\ell}(x_{N_\ell})\big)$

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

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 $\left(\mathcal{M}(\mathbb{R}), \mathrm{D}^2\right)$

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

Representer theorem for deep neural networks

Theorem ($\mathrm{TV}^{(2)}$-optimality of deep spline networks) (U. 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 $\boldsymbol{\sigma}_\ell = \left(\sigma_{1,\ell}, \dots, \sigma_{N_\ell,\ell}\right) : \mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$ 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 $b_{1,n,\ell}, b_{2,n,\ell}, a_{1,n,\ell}, \dots, a_{K_{n,\ell},n,\ell} \in \mathbb{R}$.

Deep spline networks: Discussion

- Global optimality achieved with **spline activations**

- Justification of popular schemes / Backward compatibility

  - Standard ReLU networks $(K_{n,\ell} = 1,\ \boldsymbol{b}_{n,\ell} = \boldsymbol{0})$ (Glorot *ICAIS* 2011) (LeCun-Bengio-Hinton *Nature* 2015)

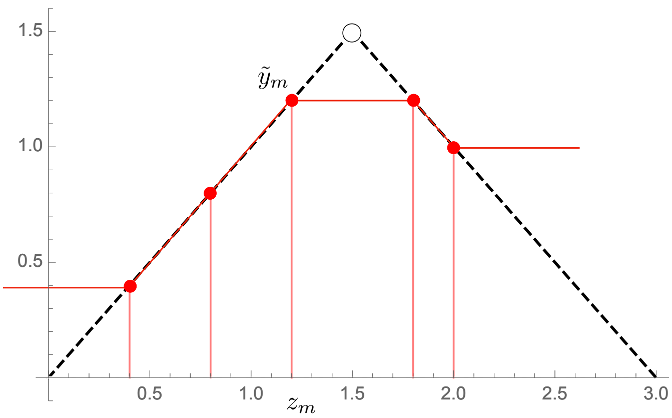
  - Linear regression: $\lambda \to \infty \Rightarrow K_{n,\ell} = 0$

  - State-of-the-art Parametric ReLU networks $(K_{n,\ell} = 1)$ (He et al. CVPR 2015)
    1 ReLU + linear term (per neuron)

  - Adaptive-piecewise linear (APL) networks $(K_{n,\ell} = 5 \text{ or } 7,\ \boldsymbol{b}_{n,\ell} = \boldsymbol{0})$ (Agostinelli et al. 2015)

Comparison of linear interpolators

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



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

Conclusion

- Unifying result that supports all known "representer" theorems
  - Classical methods based on **quadratic minimization**
    - Kernel-based methods for RKHS (Poggio-Girosi 1990; Schölkopf 2001)
    - Tikhonov regularization (Tikhonov 1977; Gupta 2018)
  - Optimization in **reflexive and strictly-convex** Banach spaces
    - L_p splines (de Boor 1976; ...)
    - Reproducing kernel Banach spaces (Zhang-Xu-Zhang 2009; Zhang-Zhang 2012)
  - Modern **sparsity-based** optimization
    - ℓ_1-minimization for compressed sensing (Donoho 2006; Candes 2006; Baraniuk 2007)
    - Total variation minimization for the recovery of spikes (Candes Fernandez-Grada 2013; Duval-Peyré 2015)
    - L-splines are optimum solutions for inverse problems with generalized total-variation regularization (Unser-Fageot-Ward 2017; Flinth-Weiss 2018; Bredies-Carioni 2020)
    - Optimality of deep ReLU networks (Unser 2019)

Conclusion (Cont'd)

- Remarkable level of generality $\Rightarrow$ opens up new perspectives
  - Fundamental ingredients for applicability
    - **Banach space** that is matched to the problem at hand
    - Knowledge of **dual mapping** *vs.* **extreme points**

No need for Fréchet derivatives or sub-gradients !!!

- Sparse kernel expansions: Open computational challenge

  Efficient algorithm for **displacing**/**removing** kernels

Acknowledgments

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

- Dr. Julien Fageot
- Prof. John Paul Ward
- Harshit Gupta
- Shayan Aziznejad
- Thomas Debarre
- Dr. Emrah Bostan
- Prof. Ulugbek Kamilov
- Prof. Matthieu Guerquin-Kern

  ....

FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

and collaborators ...

- Prof. Demetri Psaltis
- Prof. Marco Stampanoni
- Prof. Carlos-Oscar Sorzano
- Dr. Arne Seitz
- ....

Selected references

- Reference books

  - A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer, 2004.
  - W. Rudin, *Functional Analysis*, McGraw-Hill Book Co., 1991.
  - L. Schwartz. *Théorie des Distributions*. Hermann, Paris, 1966.

- Selected papers

  - M. Unser, "A unifying representer theorem for inverse problems and machine learning," arXiv:1903.00687 (2019).
  - M. Unser, J. Fageot, and H. Gupta, Representer theorems for sparsity-promoting ℓ_1 regularization, *IEEE Trans. Information Theory*, 62 (2016), pp. 5167–5180.
  - M. Unser, J. Fageot, and J. P. Ward, "Splines are universal solutions of linear inverse problems with generalized-TV regularization," *SIAM Review*, 59 (2017), pp. 769–793.
  - H. Gupta, J. Fageot, and M. Unser, "Continuous-domain solutions of linear inverse problems with Tikhonov versus generalized TV regularization," *IEEE Trans. Signal Processing*, 66 (2018), pp. 4670–4684.
  - C. Boyer, A. Chambolle, Y. De Castro, V. Duval, F. De Gournay, and P. Weiss, "On representer theorems and convex regularization", *SIAM Journal of Optimization*, 29 (2019) pp. 1260-1281.
  - S. Aziznejad and M. Unser, "Multiple-kernel regression with sparsity constraint", arXiv:1811.00836, 2018.
  - M. Unser, "A representer theorem for deep neural networks," *J. Machine Learning Research*, 20 (2019), no. 110, pp. 1-30.

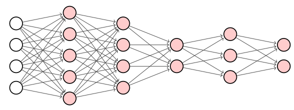
- Preprints and demos: http://bigwww.epfl.ch/

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

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

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

Tikhonov regularization (Exact solution)

$$f_0 = \arg\min_{f\in\mathcal{H}}\left(\sum_{m=1}^{M}|y_m - \langle \nu_m, f\rangle|^2 + \lambda\|f\|_{\mathcal{H}}^2\right)$$

and

$$f_0 = \mathrm{span}\{\varphi_m\}_{m=1}^{M}$$

- Equivalent finite-dimensional problem

$$\boldsymbol{a}_0 = \arg\min_{\boldsymbol{a}\in\mathbb{R}^M}\left(\|\boldsymbol{y} - \mathbf{H}\boldsymbol{a}\|^2 + \lambda \boldsymbol{a}^T\mathbf{H}\boldsymbol{a}\right)$$

- Closed-form solution

$$\boldsymbol{a}_0 = (\mathbf{H}\mathbf{H} \ + \lambda\mathbf{H})^{-1}\mathbf{H}\boldsymbol{y} = (\mathbf{H} + \lambda\mathbf{I})^{-1}\boldsymbol{y}$$

$\mathbf{H}$ invertible $\quad\Leftrightarrow\quad$ ν_m are linearly independent