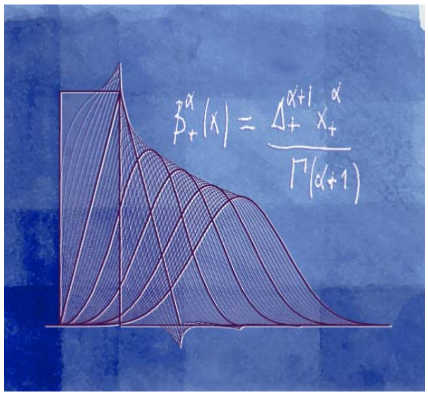
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

New representer theorems: From compressed sensing to deep learning

Michael Unser
Biomedical Imaging Group
EPFL, Lausanne, Switzerland

Joint work with
Julien Fageot, John-Paul Ward
and Kyong Jin



Mathematisches Kolloquium, Universität Wien, October 24, 2018, Wien, Austria

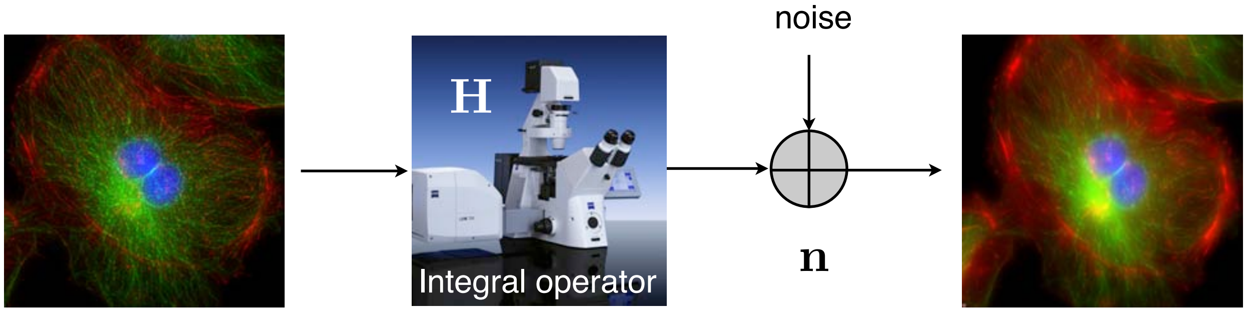
OUTLINE

- **Introduction**
  - Image reconstruction as an inverse problem
  - Learning as an inverse problem
- **Prologue: discrete-domain regularization**
- **Continuous-domain theory**
  - Splines and operators
  - L_2 regularization (theory of RKHS) : classical representer theorem
  - gTV regularization: representer theorem for CS
- **From compressed sensing to deep networks**
  - Unrolling forward/backward iterations: FBPConv
  - New representer theorem for deep neural networks

FNSNF
Swiss National Science Foundation
erc²

Variational formulation of inverse problem

- Linear forward model $\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}$



$\mathbf{s}$

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

- Reconstruction as an optimization problem

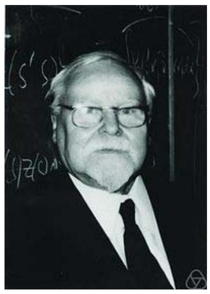
$$\mathbf{s}_{\mathrm{rec}} = \arg\min_{\mathbf{s}\in\mathbb{R}^N} \underbrace{\|\mathbf{y} - \mathbf{H}\mathbf{s}\|_2^2}_{\text{data consistency}} + \underbrace{\lambda\|\mathbf{L}\mathbf{s}\|_p^p}_{\text{regularization}}, \quad p = 1, 2$$

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)



Andrey N. Tikhonov (1906-1993)

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

- 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

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 $f : \mathbb{R}^N \to \mathbb{R}$ such that $f(\boldsymbol{x}_m) \approx y_m$ 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

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

- Regularized least-squares fit

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

Unifying continuous-domain formulation

Unknown is a function $f : \mathbb{R}^d \to \mathbb{R}$

- Regularization functional: $R(f) : \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}^+$

  Promotes smoothness (Sobolev norm) or sparsity (gTV)

- Native space $\mathcal{B}(\mathbb{R}^d)$ ***Banach*** *vs. Hilbert space (RKHS)*

$$\mathcal{B}(\mathbb{R}^d) = \{f : \mathbb{R}^d \to \mathbb{R} : R(f) < \infty\}$$

- Linear measurement operator $\mathbf{H} : \mathcal{B} \to \mathbb{R}^M$ ***Linear functionals*** *vs. point values*

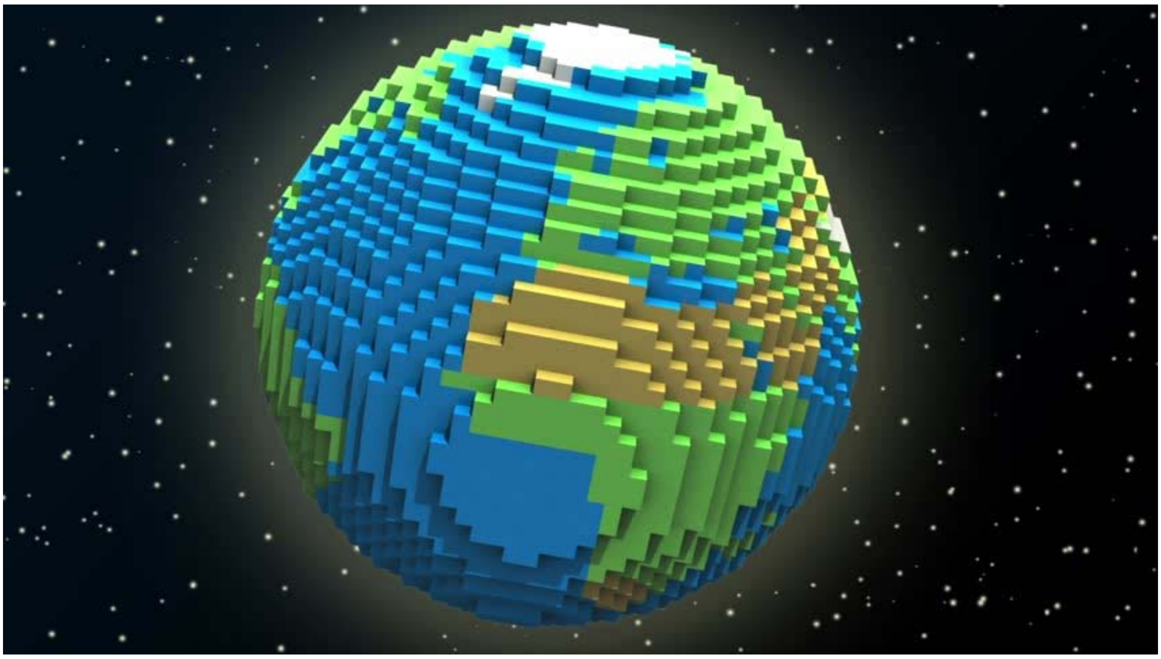
$$\mathbf{H} = (h_1, \dots, h_M) : f \mapsto \big(\langle h_1, f\rangle, \dots, \langle h_M, f\rangle\big)$$

- Regularized functional fit to the data ***Arbitrary convex loss*** *vs. least squares*

$$f_{\mathrm{opt}} = \arg\min_{f \in \mathcal{B}} \left(\underbrace{\sum_{m=1}^{M} \left|y_m - \langle h_m, f\rangle\right|^2}_{E(\boldsymbol{y}, \mathbf{H}\{f\})} + \lambda R(f) \right)$$

(Schölkopf 2001; Rosasco 2004)

Prologue: Discrete-domain regularization



Classical least-squares fit with l_2 regularization

- Linear measurement model:
  $y_m = \langle \mathbf{h}_m, \mathbf{x} \rangle + n[m], \quad m = 1, \dots, M$
- System matrix of size $M \times N$: $\mathbf{H} = [\mathbf{h}_1 \cdots \mathbf{h}_M]^T$

$$\mathbf{x}_{\mathrm{LS}} = \arg \min_{\mathbf{x} \in \mathbb{R}^N} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_2^2$$

$$\Rightarrow \quad \mathbf{x}_{\mathrm{LS}} = (\mathbf{H}^T\mathbf{H} + \lambda \mathbf{I}_N)^{-1}\mathbf{H}^T\mathbf{y}$$

$$= \mathbf{H}^T\mathbf{a} = \sum_{m=1}^{M} a_m \mathbf{h}_m \quad \text{where} \quad \mathbf{a} = (\mathbf{H}\mathbf{H}^T + \lambda \mathbf{I}_M)^{-1}\mathbf{y}$$

Interpretation: $\mathbf{x}_{\mathrm{LS}} \in \mathrm{span}\{\mathbf{h}_m\}_{m=1}^{M}$

Lemma
$(\mathbf{H}^T\mathbf{H} + \lambda \mathbf{I}_N)^{-1}\mathbf{H}^T = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T + \lambda \mathbf{I}_M)^{-1}$

Switch to l_1 regularization ⇒ sparsifying effect

- Linear measurement model:
  $y_m = \langle \mathbf{h}_m, \mathbf{x} \rangle + n[m], \quad m = 1, \dots, M$
- System matrix of size $M \times N$: $\mathbf{H} = [\mathbf{h}_1 \cdots \mathbf{h}_M]^T$

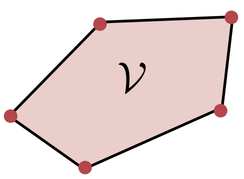
$$\text{(P1):} \quad \mathcal{V} = \arg \min_{\mathbf{x} \in \mathbb{R}^N} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_{\ell_1}$$

Representer theorem for unconstrained ℓ_1 minimization

The solution set $\mathcal{V}$ of (P1) is convex, compact with extreme points of the form

$$\mathbf{x}_{\text{sparse}} = \sum_{k=1}^{K} a_k \mathbf{e}_{n_k} \quad \text{with} \quad K = \|\mathbf{x}_{\text{sparse}}\|_0 \leq M.$$

element of canonical basis with $[\mathbf{e}_n]_m = \delta_{m-n}$



If CS condition on $\mathbf{H}$ is satisfied, then solution is unique

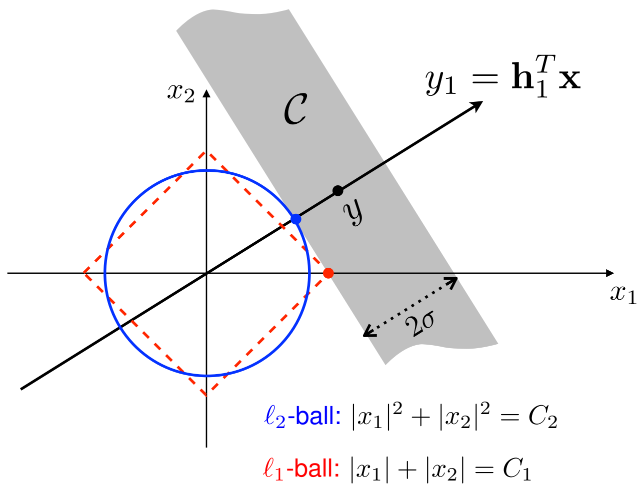
(U.-Fageot-Gupta *IEEE Trans. Info. Theory*, Sept. 2016)

Geometry of l_2 vs. l_1 minimization

- Prototypical inverse problem

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

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

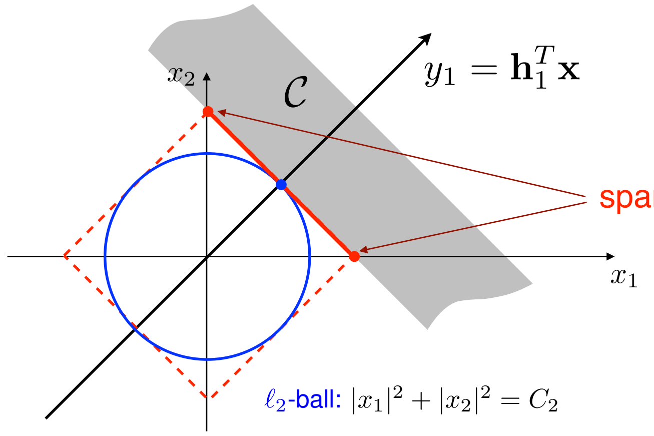


Geometry of l_2 vs. l_1 minimization

- 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

Part II: Continuous-domain theory

Continuous-domain regularization (L_2 scenario)

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

L: suitable differential operator

- Theory of reproducing kernel Hilbert spaces (Aronszajn 1950)

  $\langle f, g\rangle_{\mathcal{H}} = \langle \mathrm{L}f, \mathrm{L}g\rangle$

- Interpolation and approximation theory
  - Smoothing splines (Schoenberg 1964, Kimeldorf-Wahba 1971)
  - Thin-plate splines, radial basis functions (Duchon 1977)
- Machine learning
  - Radial basis functions, kernel methods (Poggio-Girosi 1990)
  - Representer theorem(s) (Schölkopf-Smola 2001)

Splines are analog, but intrinsically sparse

$\mathrm{L}\{\cdot\}$: admissible differential operator

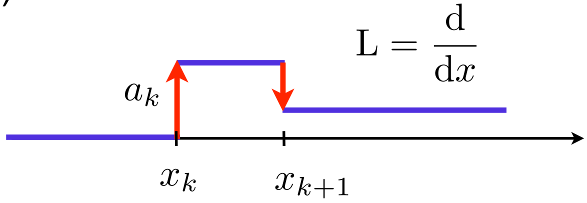
$\delta(\cdot - \boldsymbol{x}_0)$: Dirac impulse shifted by $\boldsymbol{x}_0 \in \mathbb{R}^d$

Definition

The function $s : \mathbb{R}^d \to \mathbb{R}$ is a (non-uniform) L-spline with knots $(\boldsymbol{x}_k)_{k=1}^K$ if

$$\mathrm{L}\{s\} = \sum_{k=1}^{K} a_k \delta(\cdot - \boldsymbol{x}_k) = w_\delta \quad : \text{ spline's innovation}$$

Spline theory: (Schultz-Varga, 1967)

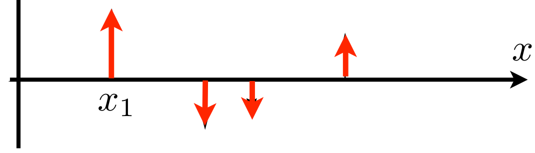


Spline synthesis: example

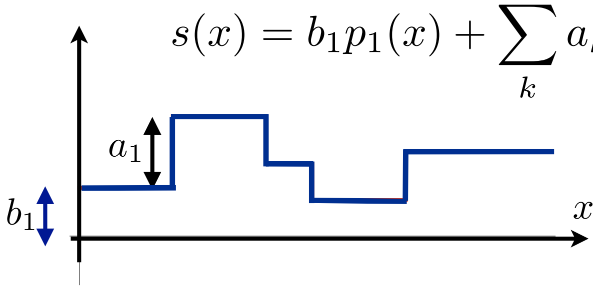
$\mathrm{L} = \mathrm{D} = \dfrac{\mathrm{d}}{\mathrm{d}x}$ Null space: $\mathcal{N}_{\mathrm{D}} = \mathrm{span}\{p_1\}, \quad p_1(x) = 1$

$\rho_{\mathrm{D}}(x) = \mathrm{D}^{-1}\{\delta\}(x) = \mathbb{1}_+(x)$: Heaviside function

$$w_\delta(x) = \sum_k a_k \delta(x - x_k)$$



$$s(x) = b_1 p_1(x) + \sum_k a_k \mathbb{1}_+(x - x_k)$$



Spline synthesis: generalization

L: spline admissible operator (LSI)

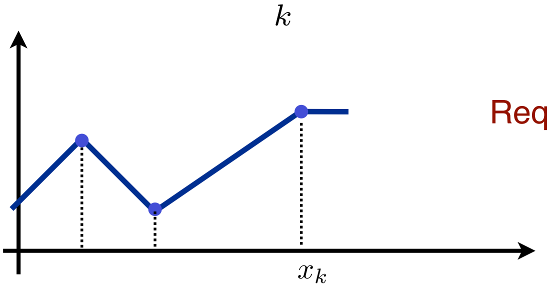
$\rho_{\mathrm{L}}(\boldsymbol{x}) = \mathrm{L}^{-1}\{\delta\}(\boldsymbol{x})$: Green's function of L

Finite-dimensional null space: $\mathcal{N}_{\mathrm{L}} = \mathrm{span}\{p_n\}_{n=1}^{N_0}$

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

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

Requires specification of boundary conditions



RKHS representer theorem for L_2 regularization

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

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

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

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

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

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

(Schölkopf-Smola 2001)

Representer theorem for L_2-regularization

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

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

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

L_2 representer theorem for variational splines

Theoretical difficulty: $\|f\|^2_{\mathcal{H}} \longrightarrow \|\mathrm{L}f\|^2_{L_2}$ (only a semi-norm !)

(P2) $$\arg\min_{f\in\mathcal{H}_{\mathrm{L}}}\left(\sum_{m=1}^{M}|y_m - f(\boldsymbol{x}_m)|^2 + \lambda\|\mathrm{L}f\|^2_{L_2(\mathbb{R}^d)}\right)$$

$\rho_{\mathrm{L}^*\mathrm{L}}(\boldsymbol{x}) = (\mathrm{L}^*\mathrm{L})^{-1}\{\delta\}(\boldsymbol{x})$: **Green's function** of $(\mathrm{L}^*\mathrm{L})$

(Schoenberg 1964, Kimeldorf-Wahba 1971)

L_2 representer theorem for variational splines

The solution of (P2) is unique and of the form

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} a_m \rho_{\mathrm{L}^*\mathrm{L}}(\boldsymbol{x}-\boldsymbol{x}_m) \quad + \quad \sum_{n=1}^{N_0} b_n p_n(\boldsymbol{x});$$

i.e., it is a $(\mathrm{L}^*\mathrm{L})$-spline with knots at the $\{\boldsymbol{x}_m\}$.

Example: $\mathrm{L} = \mathrm{D}^2$ with $\rho_{\mathrm{D}^4}(x) \propto |x|^3 \quad\Rightarrow\quad f(x)$ is a cubic spline

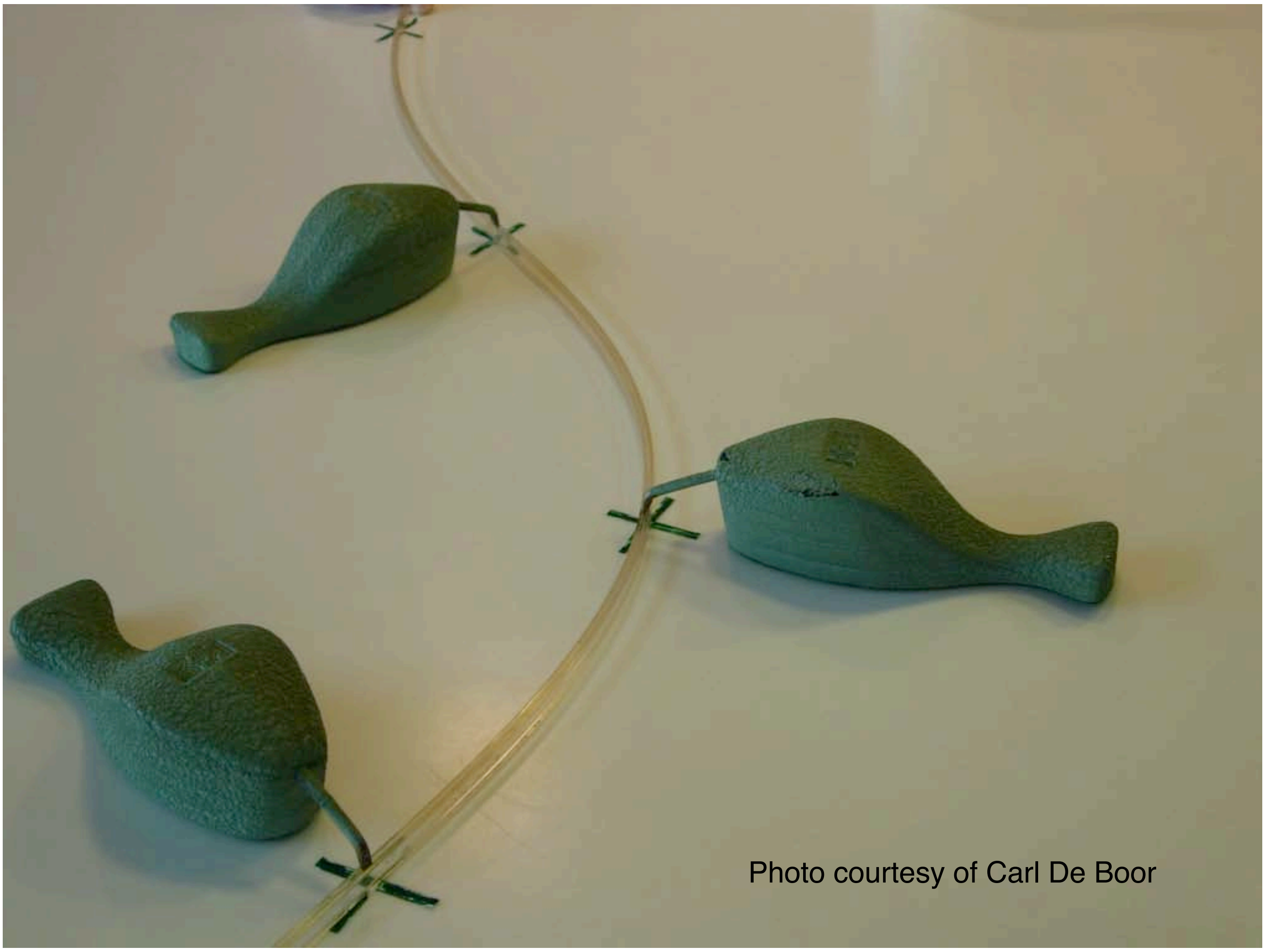


Photo courtesy of Carl De Boor

Quest for sparsity in a continuous world

Sparsity and continuous-domain modeling

- Compressed sensing (CS)
  - Generalized sampling and infinite-dimensional CS (Adcock-Hansen 2011)
  - Xampling: CS of analog signals (Eldar 2011)
  - Recovery of Dirac impulses from Fourier measurements (Vetterli et al. 2002)
    (Bredies 2013; Candes & Fernandez-Granda 2014; Duval-Peyré 2015)
- Splines and approximation theory
  - L_1 splines (Fisher-Jerome 1975)
  - Locally-adaptive regression splines (Mammen-van de Geer 1997)
  - Generalized TV (Steidl et al. 2005; Bredies et al. 2010)
- Statistical modeling
  - Sparse stochastic processes (Unser et al. 2011-2014)

Proper continuous counterpart of $\ell_1(\mathbb{Z}^d)$

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

$\mathcal{S}'(\mathbb{R}^d)$: Schwartz's space of tempered distributions

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

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

where $w : \varphi \mapsto \langle w, \varphi\rangle = \int_{\mathbb{R}^d} \varphi(\boldsymbol{r}) w(\boldsymbol{r}) \mathrm{d}\boldsymbol{r}$

- Equivalent definition of “total variation” norm

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

- Basic inclusions
  - $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{M}(\mathbb{R}^d)$ with $\|\delta(\cdot - \boldsymbol{x}_0)\|_{\mathcal{M}} = 1$ for any $\boldsymbol{x}_0 \in \mathbb{R}^d$
  - $\|f\|_{\mathcal{M}} = \|f\|_{L_1(\mathbb{R}^d)}$ for all $f \in L_1(\mathbb{R}^d) \quad \Rightarrow \quad L_1(\mathbb{R}^d) \subseteq \mathcal{M}(\mathbb{R}^d)$

Representer theorem for gTV regularization

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

- 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)

Convex loss function: $F : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ $\qquad \boldsymbol{\nu} : \mathcal{M}_{\mathrm{L}} \to \mathbb{R}^M$

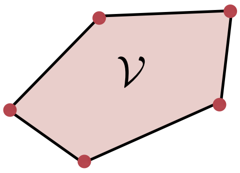
(P1') $\arg \min_{f \in \mathcal{M}_{\mathrm{L}}(\mathbb{R}^d)} \big(F(\boldsymbol{y}, \boldsymbol{\nu}(f)) + \lambda \|\mathrm{L} f\|_{\mathcal{M}} \big)$ with $\boldsymbol{\nu}(f) = \big(\langle h_1, f \rangle, \dots, \langle h_M, f \rangle\big)$

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

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

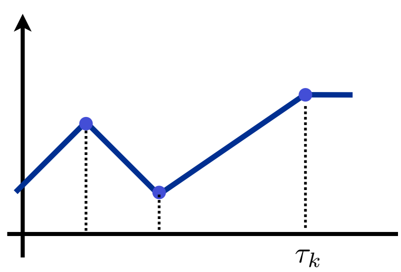
- Total 2nd-variation: $\mathrm{TV}^{(2)}(s) = \sup_{\|\varphi\|_\infty \le 1} \langle \mathrm{D}^2 s, \varphi \rangle = \|\mathrm{D}^2 s\|_{\mathcal{M}}$

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

- Generic form of the solution

$$s_{\mathrm{spline}}(x) = b_1 + b_2 x + \sum_{k=1}^{K} a_k (x - \tau_k)_+$$

no penalty



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)

  $\Rightarrow$ polynomial splines of degree $(n-1)$ (Schoenberg 1946)

- $\mathrm{L} = \mathrm{D}^n + a_{n-1}\mathrm{D}^{n-1} + \cdots + a_0\mathrm{I}$ (ordinary differential operator)

  $\Rightarrow$ exponential splines (Dahmen-Micchelli 1987)

- Fractional derivatives: $\mathrm{L} = \mathrm{D}^\gamma \quad \xleftrightarrow{\mathcal{F}} \quad (\mathrm{j}\omega)^\gamma$

  $\Rightarrow$ fractional splines (U.-Blu 2000)

- Fractional Laplacian: $(-\Delta)^{\frac{\gamma}{2}} \quad \xleftrightarrow{\mathcal{F}} \quad \|\boldsymbol{\omega}\|^\gamma$

  $\Rightarrow$ polyharmonic splines (Duchon 1977)

- Elliptical differential operators; e.g, $\mathrm{L} = (-\Delta + \alpha\mathrm{I})^\gamma$

  $\Rightarrow$ Sobolev splines (Ward-U. 2014)

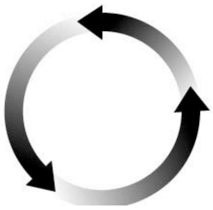
Discretization: compatible with CS paradigm

$$\mathbf{s}_{\mathrm{sparse}} = \arg\min_{\mathbf{s}\in\mathbb{R}^K} \left(\frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{s}\|_2^2 + \lambda\|\mathbf{u}\|_1\right) \text{ subject to } \mathbf{u} = \mathbf{L}\mathbf{s}$$

ADMM algorithm

$$\mathcal{L}_{\mathcal{A}}(\mathbf{s}, \mathbf{u}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{s}\|_2^2 + \lambda\sum_n |[\mathbf{u}]_n| + \boldsymbol{\alpha}^T(\mathbf{L}\mathbf{s} - \mathbf{u}) + \frac{\mu}{2}\|\mathbf{L}\mathbf{s} - \mathbf{u}\|_2^2$$

For $k = 0, \ldots, K$



Linear step

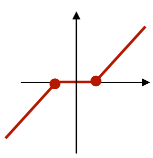
$$\mathbf{s}^{k+1} = \left(\mathbf{H}^T\mathbf{H} + \mu\mathbf{L}^T\mathbf{L}\right)^{-1}\left(\mathbf{z}_0 + \mathbf{z}^{k+1}\right)$$

with $\mathbf{z}^{k+1} = \mathbf{L}^T\left(\mu\mathbf{u}^k - \boldsymbol{\alpha}^k\right)$

$$\boldsymbol{\alpha}^{k+1} = \boldsymbol{\alpha}^k + \mu\left(\mathbf{L}\mathbf{s}^{k+1} - \mathbf{u}^k\right)$$

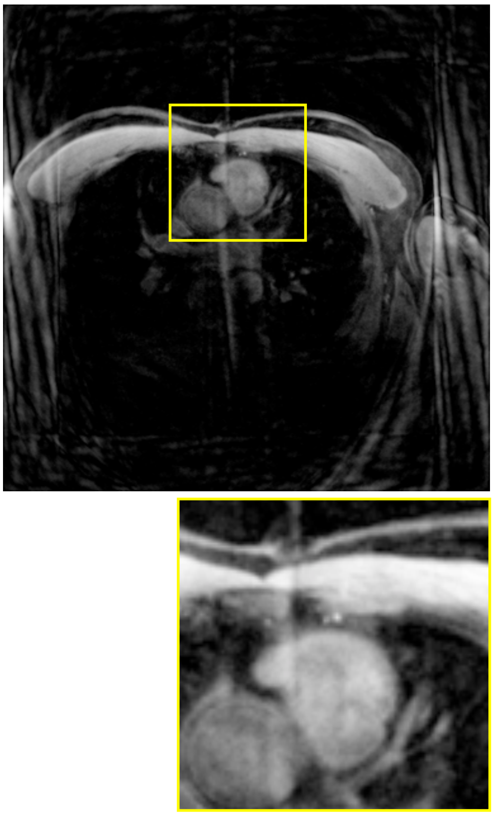
Proximal step = pointwise non-linearity

$$\mathbf{u}^{k+1} = \mathrm{prox}_{|\cdot|}\left(\mathbf{L}\mathbf{s}^{k+1} + \tfrac{1}{\mu}\boldsymbol{\alpha}^{k+1}; \tfrac{\sigma^2}{\mu}\right)$$

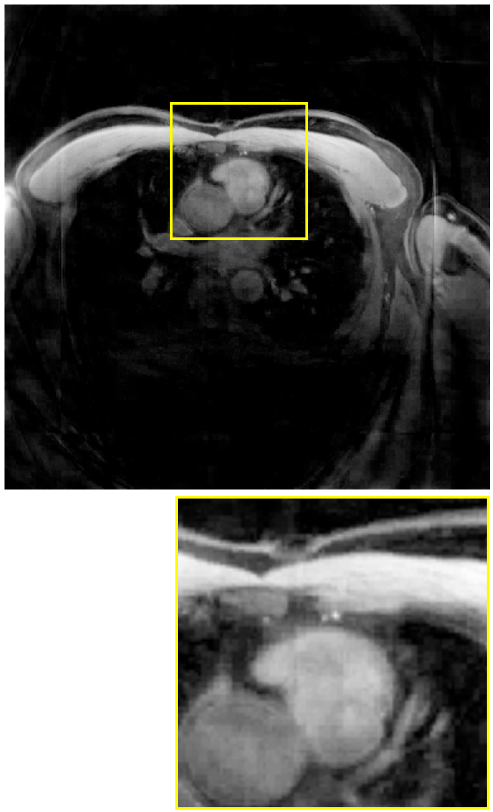


Example: ISMRM reconstruction challenge

L_2 regularization (Laplacian)



ℓ_1 / TV regularization



(Guerquin-Kern *IEEE TMI* 2011)

OUTLINE

- Linear inverse problems and regularization ✔
- Continuous-domain theory ✔
  - Splines and operators
  - Classical L_2 regularization: theory of RKHS
  - Minimization of gTV: the optimality of splines
- **From compressed sensing to deep networks**
  - Image recovery with sparsity constraints
  - FBPConvNet
  - Representer theorem for deep neural networks

When is unrolled ADMM a deep ConvNet ?

Answer: when $\mathbf{H}^T\mathbf{H}$ and $\mathbf{L}$ are both convolutions

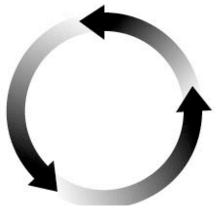
$$\mathcal{L}_{\mathcal{A}}(\mathbf{s},\mathbf{u},\boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{y}-\mathbf{H}\mathbf{s}\|_2^2 + \sigma^2\sum_n |[\mathbf{u}]_n| + \boldsymbol{\alpha}^T(\mathbf{L}\mathbf{s}-\mathbf{u}) + \frac{\mu}{2}\|\mathbf{L}\mathbf{s}-\mathbf{u}\|_2^2$$

ADMM algorithm

Initialization

$$z_0 = \mathbf{H}^T\mathbf{y} \qquad \mathbf{u}^0 = \mathbf{0}$$
$$\mathbf{s}^0 = \mathbf{0} \qquad \boldsymbol{\alpha}^0 = \mathbf{0}$$

For $k = 0, \dots, K$



Linear step = Convolutions

$$\mathbf{s}^{k+1} = \left(\mathbf{H}^T\mathbf{H} + \mu\mathbf{L}^T\mathbf{L}\right)^{-1}\left(\mathbf{z}_0 + \mathbf{z}^{k+1}\right)$$

with $\mathbf{z}^{k+1} = \mathbf{L}^T\left(\mu\mathbf{u}^k - \boldsymbol{\alpha}^k\right)$

$$\boldsymbol{\alpha}^{k+1} = \boldsymbol{\alpha}^k + \mu\left(\mathbf{L}\mathbf{s}^{k+1} - \mathbf{u}^k\right)$$

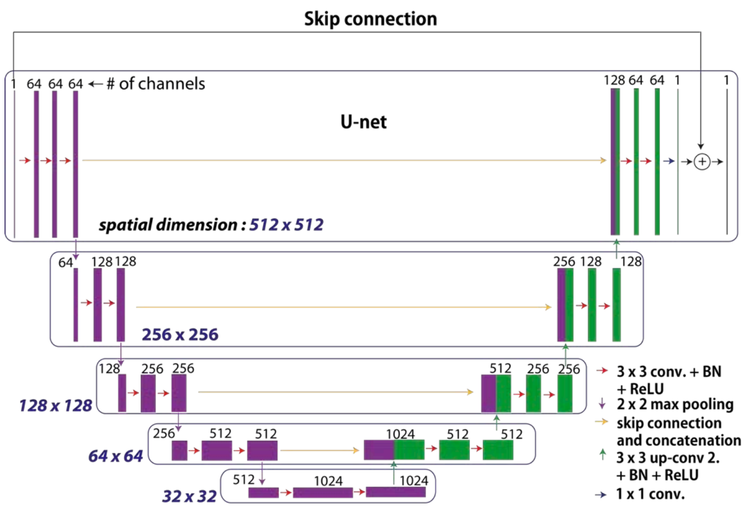
Proximal step = pointwise non-linearity

$$\mathbf{u}^{k+1} = \mathrm{prox}_{|\cdot|}\left(\mathbf{L}\mathbf{s}^{k+1} + \tfrac{1}{\mu}\boldsymbol{\alpha}^{k+1}; \tfrac{\sigma^2}{\mu}\right)$$

Recent appearance of Deep ConvNets

(Jin et al. 2016; Adler-Öktem 2017; Chen et al. 2017; ...)

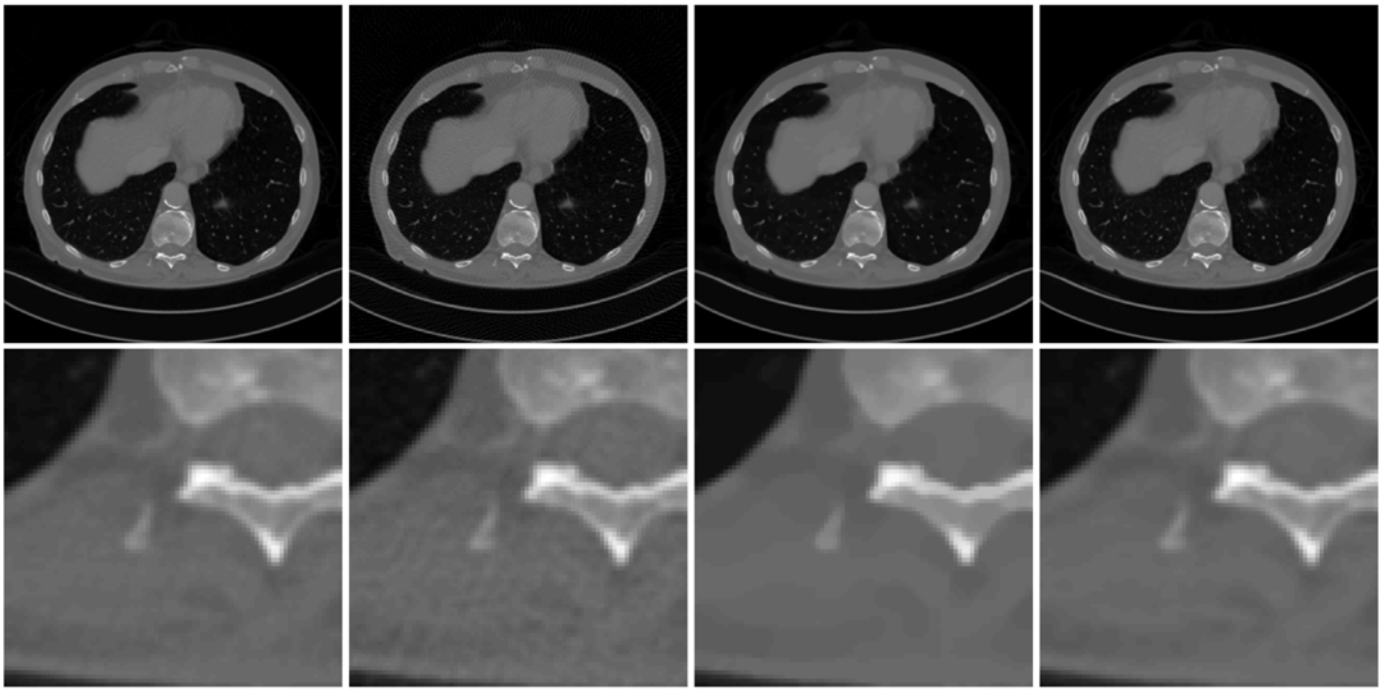
- CT reconstruction based on Deep ConvNets
  - Input: Sparse view FBP reconstruction
  - Training: Set of 500 high-quality full-view CT reconstructions
  - Architecture: U-Net with skip connection (Jin et al., IEEE TIP 2017)



CT data

Dose reduction by 7: 143 views

Ground truth	FBP SNR 24.06	TV SNR 29.64	FBPConvNet SNR 35.38



Reconstructed from
from 1000 views

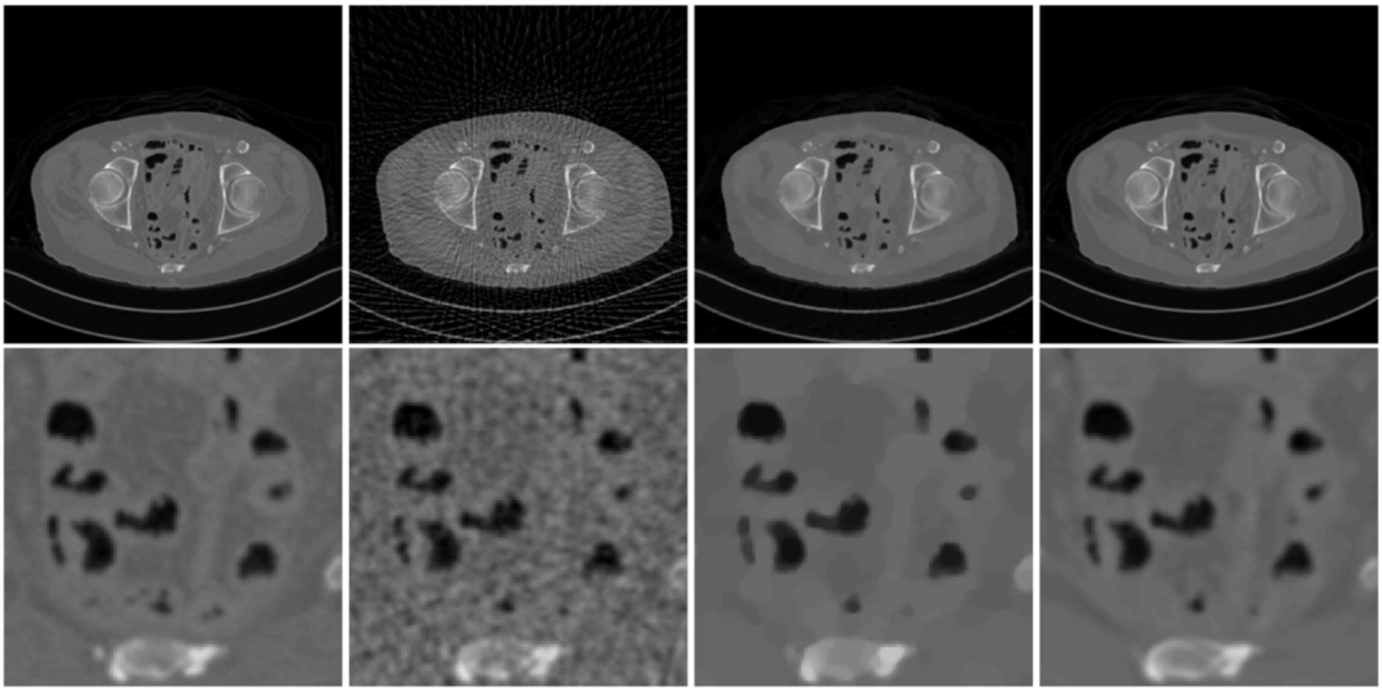
(Jin et al., *IEEE Trans. Im Proc.*, 2017)

MAYO CLINIC

CT data

Dose reduction by 20: 50 views

Ground truth	FBP SNR 13.43	TV SNR 24.89	FBPConvNet SNR 28.53



Reconstructed from
from 1000 views

(Jin et al., *IEEE Trans. Im Proc.*, 2017)

MAYO CLINIC

Finale:

Representer theorem for deep learning

Deep neural networks and splines

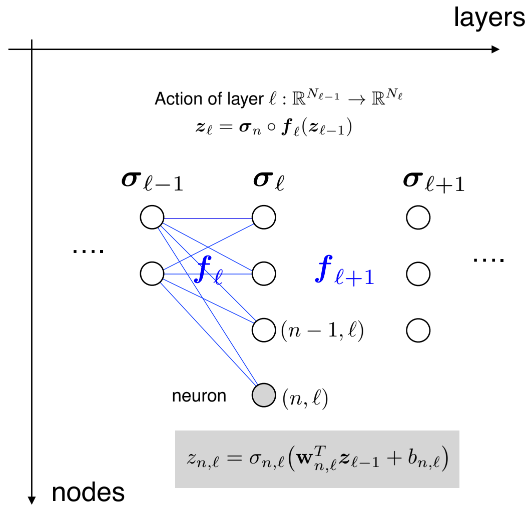
- Preferred choice of activation function: ReLU
  - ReLU works nicely with dropout / ℓ_1-regularization (Glorot *ICAIS* 2011)
  - Networks with hidden ReLU are easier to train
  - State-of-the-art performance (LeCun-Bengio-Hinton *Nature* 2015)

- Deep nets as Continuous PieceWise-Linear maps
  - MaxOut $\Rightarrow$ CPWL (Goodfellow *PMLR* 2013)
  - ReLU $\Rightarrow$ CPWL (Montufar *NIPS* 2014)
  - CPWL $\Rightarrow$ Deep ReLU network (Wang-Sun *IEEE-IT* 2005)

Feedforward deep neural network

- Layers: $\ell = 1, \dots, L$
- Deep structure descriptor: $(N_0, N_1, \cdots, N_L)$
- Neuron or node index: $(n, \ell), \quad n = 1, \cdots, N_\ell$
- Activations functions: $\sigma_{n,\ell} : \mathbb{R} \to \mathbb{R}$

- 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: $\mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$
  $\boldsymbol{\sigma}_\ell : \boldsymbol{x} \mapsto \boldsymbol{\sigma}_\ell(\boldsymbol{x}) = \big(\sigma_{1,\ell}(x_1), \dots, \sigma_{N_\ell,\ell}(x_{N_\ell})\big)$



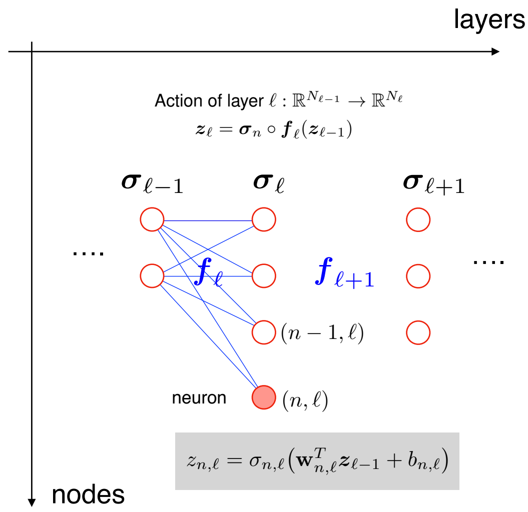
Conventional design: $\sigma_{n,\ell} = \sigma$

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

Deep neural net with optimized activations

- Layers: $\ell = 1, \dots, L$
- Deep structure descriptor: $(N_0, N_1, \cdots, N_L)$
- Neuron or node index: $(n, \ell), \quad n = 1, \cdots, N_\ell$
- Activations functions: $\sigma_{n,\ell} : \mathbb{R} \to \mathbb{R}$

- 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: $\mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$
  $\boldsymbol{\sigma}_\ell : \boldsymbol{x} \mapsto \boldsymbol{\sigma}_\ell(\boldsymbol{x}) = \big(\sigma_{1,\ell}(x_1), \dots, \sigma_{N_\ell,\ell}(x_{N_\ell})\big)$



New **adaptive** design: $x \mapsto \sigma_{n,\ell}(x)$ s.t. $\mathrm{TV}^{(2)}(\sigma_{n,\ell})$ minimum

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

New representer theorem for deep neural networks

(Unser, *arXiv:1802.09210*, Feb 2018)

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

- 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 of M data points $\boldsymbol{y}_m \approx \mathbf{f}(\boldsymbol{x}_m)$, 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
- $R_\ell : \mathbb{R}^{N_\ell \times N_\ell} \to \mathbb{R}^+$: 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

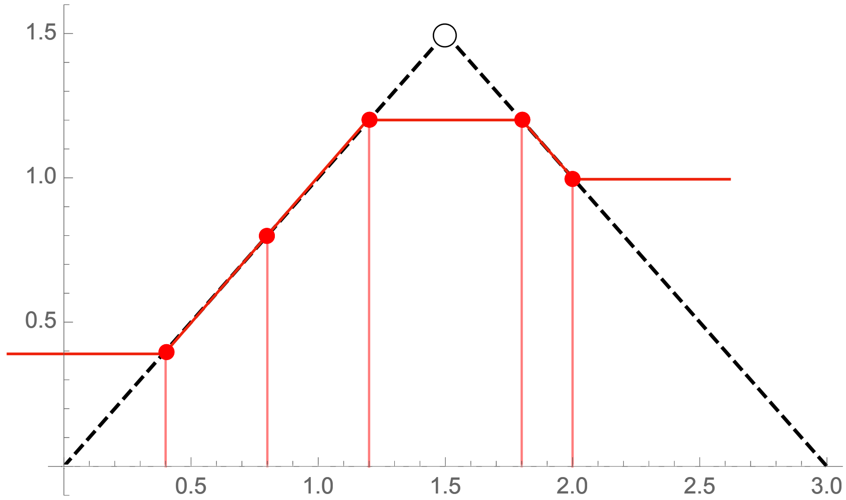
- its number $0 \le K = K_{n,\ell}$ of knots (ideally, much smaller than M);
- the location $\{\tau_k = \tau_{k,n,\ell}\}_{k=1}^{K_{n,\ell}}$ of these knots (ReLU biases);
- the expansion coefficients $\mathbf{b}_{n,\ell} = (b_{1,n,\ell}, b_{2,n,\ell}) \in \mathbb{R}^2$, $\boldsymbol{a}_{n,\ell} = (a_{1,n,\ell},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

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

- Link with ℓ_1 minimization techniques

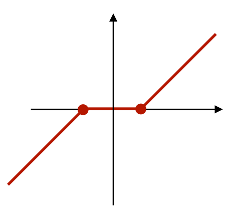
$$\mathrm{TV}^{(2)}\{\sigma_{n,\ell}\} = \sum_{k=1}^{K_{n,\ell}} |a_{k,n,\ell}| = \|\mathbf{a}_{n,\ell}\|_1$$

Comparison of linear interpolators



Deep spline networks: Discussion

- Global optimality achieved with **spline activations**
- State-of-the-art ReLU networks $(K_{n,\ell} = 1,\ \boldsymbol{b}_{n,\ell} = \mathbf{0})$
  - No need to normalize:
    $(\mathbf{w}_{n,\ell}^T \boldsymbol{x} - z_{n,\ell})_+ = (a_{n,\ell}\mathbf{u}_{n,\ell}^T \boldsymbol{x} - z_{n,\ell})_+ = a_{n,\ell}(\mathbf{u}_{n,\ell}^T \boldsymbol{x} - \tau_{n,\ell})_+$
- Key features
  - Produces a global mapping $\boldsymbol{x} \mapsto \boldsymbol{f}(\boldsymbol{x})$ that is **continuous** and **piecewise-linear**
  - Direct control of complexity (number of knots): adjustment of λ
  - Ability to suppress unnecessary layers
- Backward compatibility
  - Linear regression: $\lambda \to \infty \Rightarrow K_{n,\ell} = 0$
  - Compressed sensing / ℓ_1 minimization



SUMMARY: Controlling smoothness **vs. sparsity**

- New findings resonate with what is known in discrete setting
  - l_2 solution lives in a **fixed** subspace of dimension M
  - Tikhonov solution is intrinsically “**blurred**”
  - Minimization of l_1 favors sparse solutions (independently of sensing matrix)

- Specificities of continuous-domain formulation $\quad s \in \mathcal{X}$
  - Functional model: class of signals + physics $\quad s \mapsto \mathbf{z} = \mathrm{H}\{s\}$
  - Smoothing-splines: minimum “spline” energy $\quad (\mathrm{L}^*\mathrm{L})\{s_{\mathrm{smooth}}\} = \sum_{m=1}^{M} a_m h_m$
  - L-splines = signals with “sparsest” innovation $\quad \mathrm{L}\{s_{\mathrm{sparse}}\} = \sum_{k=1}^{K} a_k \delta(\cdot - \boldsymbol{x}_k)$

- Practical implications
  - Infinite-dimensional optimization is feasible (parametric form of solution)
  - gTV regularization favors **sparse** innovations with **adaptive** knots
  - Non-uniform L-splines: ***universal*** solutions of linear inverse problems

and deep neural networks …

Acknowledgments

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

- Dr. Pouya Tafti
- Prof. Arash Amini
- Dr. Emrah Bostan
- Dr. Masih Nilchian
- Dr. Ulugbek Kamilov
- Dr. Cédric Vonesch
- ....

and collaborators ...

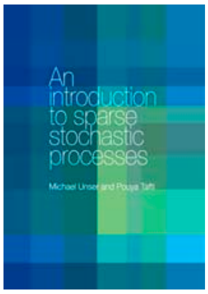
- Prof. Demetri Psaltis
- Prof. Marco Stampanoni
- Prof. Carlos-Oscar Sorzano
- Dr. Arne Seitz
- ....

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

References

Theoretical results on sparsity-promoting regularization

- M. Unser, J. Fageot, H. Gupta, “Representer Theorems for Sparsity-Promoting ℓ_1 Regularization,” *IEEE Trans. Information Theory*, Vol. 62, No. 9, pp. 5167-5180.
- 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.



Theory of sparse stochastic processes

- M. Unser and P. Tafti, ***An Introduction to Sparse Stochastic Processes***, Cambridge University Press, 2014.
  Preprint, available at `http://www.sparseprocesses.org`.
- **For splines**: see chapter 6
- E. Bostan, U.S. Kamilov, M. Nilchian, M. Unser, “Sparse Stochastic Processes and Discretization of Linear Inverse Problems,” *IEEE Trans. Image Processing*, vol. 22, no. 7, pp. 2699-2710, 2013.

Deep 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.
- M.T. McCann, K.H. Jin, M. Unser, “Convolutional Neural Networks for Inverse Problems in Imaging—A Review,” *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 85-95, Nov. 2017.
- M. Unser, “A representer theorem for deep neural networks,” preprint, Feb 2018, arXiv:1802.09210.