Lausanne, August 19, 2004

Dear Dr. Liebling,

I am pleased to inform you that you were selected to receive the 2004 Research Award of the Swiss Society of Biomedical Engineering for your thesis work “On Fresnelets, interference fringes, and digital holography”. The award will be presented during the general assembly of the SSBE, September 3, Zurich, Switzerland.

Please let us know if
1) you will be present to receive the award,
2) you would be willing to give a 10 minutes presentation of the work during the general assembly.

The award comes with a cash prize of 1000.- CHF.

Would you please send your banking information to the treasurer of the SSBE, Uli Diermann (Email: uli.diermann@bfh.ch), so that he can transfer the cash prize to your account?

I congratulate you on your achievement.

With best regards,

Michael Unser, Professor
Chairman of the SSBE Award Committee

cc: Ralph Mueller, president of the SSBE; Uli Diermann, treasurer

Dr. Michael Liebling
Biological Imaging Center
California Inst. of Technology
Mail Code 139-74
Pasadena, CA 91125, USA

BIOMEDICAL IMAGING GROUP (BIG)
LABORATOIRE D’IMAGERIE BIOMEDICALE
EPFL LIB
Bât. BM 4.127
CH 1015 Lausanne
Switzerland

Téléphone : + 4121 693 51 85
Fax : + 4121 693 37 01
E-mail : Michael.Unser@epfl.ch
Site web : http://bigwww.epfl.ch

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
  - $g$TV regularization: representer theorem for CS

- From compressed sensing to deep networks
  - Unrolling forward/backward iterations: FBPCConv
  - New representer theorem for deep neural networks
Variational formulation of inverse problem

- Linear forward model
  \[ y = Hs + n \]

Problem: recover \( s \) from noisy measurements \( y \)

- Reconstruction as an optimization problem
  \[
  s_{\text{rec}} = \arg \min_{s \in \mathbb{R}^N} \left( \| y - Hs \|_2^2 + \lambda \| Ls \|_p^p \right), \quad p = 1, 2
  \]

Linear inverse problems (20th century theory)

- Dealing with **ill-posed problems**: Tikhonov regularization
  \[
  R(s) = \| Ls \|_2^2 : \text{regularization (or smoothness) functional}
  \]
  \[
  L : \text{regularization operator (i.e., Gradient)}
  \]

  \[
  \min_s R(s) \quad \text{subject to} \quad \| y - Hs \|_2^2 \leq \sigma^2
  \]

- Equivalent variational problem
  \[
  s^* = \arg \min \left( \| y - Hs \|_2^2 + \lambda \| Ls \|_2^2 \right)
  \]

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

Interpretation: “filtered” backprojection
Learning as a (linear) inverse problem
but an infinite-dimensional one ...

Given the data points \((x_m, y_m) \in \mathbb{R}^{N+1}\), find \(f : \mathbb{R}^N \to \mathbb{R}\) such that \(f(x_m) \approx y_m\) for \(m = 1, \ldots, M\)

- Introduce smoothness or regularization constraint

\[
R(f) = \|f\|_H^2 = \|Lf\|_{L^2}^2 = \int_{\mathbb{R}^N} |Lf(x)|^2 \, dx: \text{regularization functional}
\]

\[
\min_{f \in H} R(f) \quad \text{subject to} \quad \sum_{m=1}^{M} |y_m - f(x_m)|^2 \leq \sigma^2
\]

- Regularized least-squares fit

\[
f_{\text{RKHS}} = \arg \min_{f \in H} \left( \sum_{m=1}^{M} |y_m - f(x_m)|^2 + \lambda \|f\|_H^2 \right) \Rightarrow \text{kernel estimator}
\]

Unifying continuous-domain formulation

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

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

  Promotes smoothness (Sobolev norm) or sparsity (gTV)

- Native space \(B(\mathbb{R}^d)\)

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

- Linear measurement operator \(H : B \to \mathbb{R}^M\)

  Linear functionals vs. point values

  \(H = (h_1, \ldots, h_M) : f \mapsto (\langle h_1, f \rangle, \ldots, \langle h_M, f \rangle)\)

- Regularized functional fit to the data

\[
f_{\text{opt}} = \arg \min_{f \in B} \left( \sum_{m=1}^{M} |y_m - \langle h_m, f \rangle|^2 + \lambda R(f) \right)
\]

\(\Rightarrow E(y, H\{f\})\)

(Poggio-Girosi 1990)

(Banach vs. Hilbert space (RKHS))

(Schölkopf 2001; Rosasco 2004)
Classical least-squares fit with $l_2$ regularization

- Linear measurement model:
  \[ y_m = \langle h_m, x \rangle + n[m], \quad m = 1, \ldots, M \]

- System matrix of size $M \times N$:
  \[ H = [h_1 \cdots h_M]^T \]

\[
\begin{align*}
\mathbf{x}_{LS} &= \arg \min_{\mathbf{x} \in \mathbb{R}^N} \| \mathbf{y} - H \mathbf{x} \|_2^2 + \lambda \| \mathbf{x} \|_2^2 \\
\Rightarrow \quad \mathbf{x}_{LS} &= (H^T H + \lambda I_N)^{-1} H^T \mathbf{y} \\
&= H^T \mathbf{a} = \sum_{m=1}^{M} a_m h_m \quad \text{where} \quad \mathbf{a} = (H H^T + \lambda I_M)^{-1} \mathbf{y}
\end{align*}
\]

Interpretation: $\mathbf{x}_{LS} \in \text{span}\{h_m\}_{m=1}^{M}$

Lemma
\[
(H^T H + \lambda I_N)^{-1} H^T = H^T (HH^T + \lambda I_M)^{-1}
\]
Switch to $l_1$ regularization ⇒ sparsifying effect

- Linear measurement model:
  $$y_m = \langle h_m, x \rangle + n[m], \quad m = 1, \ldots, M$$

- System matrix of size $M \times N$: $H = [h_1 \cdots h_M]^T$

\[ (P1): \quad V = \arg \min_{x \in \mathbb{R}^N} \| y - Hx \|_2^2 + \lambda \| x \|_{\ell_1} \]

**Representer theorem for unconstrained $\ell_1$ minimization**

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

$$x_{\text{sparse}} = \sum_{k=1}^K a_k e_{n_k} \quad \text{with} \quad K = \| x_{\text{sparse}} \|_0 \leq M.$$  

If CS condition on $H$ is satisfied, then solution is unique


**Geometry of $l_2$ vs. $l_1$ minimization**

- Prototypical inverse problem

\[
\begin{align*}
\min_x \{ ||y - Hx||_{\ell_2}^2 + \lambda ||x||_{\ell_2}^2 \} & \iff \min_x ||x||_{\ell_2} \text{ subject to } ||y - Hx||_{\ell_2}^2 \leq \sigma^2 \\
\min_x \{ ||y - Hx||_{\ell_2}^2 + \lambda ||x||_{\ell_1} \} & \iff \min_x ||x||_{\ell_1} \text{ subject to } ||y - Hx||_{\ell_2}^2 \leq \sigma^2 
\end{align*}
\]

$\ell_2$-ball: $|x_1|^2 + |x_2|^2 = C_2$

$\ell_1$-ball: $|x_1| + |x_2| = C_1$
Geometry of $l_2$ vs. $l_1$ minimization

- Prototypical inverse problem

$$\min_x \{ \|y - Hx\|_2^2 + \lambda \|x\|_2^2 \} \iff \min_x \|x\|_\ell_2 \text{ subject to } \|y - Hx\|_2^2 \leq \sigma^2$$

$$\min_x \{ \|y - Hx\|_2^2 + \lambda \|x\|_\ell_1 \} \iff \min_x \|x\|_\ell_1 \text{ subject to } \|y - Hx\|_2^2 \leq \sigma^2$$

Configuration for non-unique $\ell_1$ solution

Part II: Continuous-domain theory
Continuous-domain regularization ($L_2$ scenario)

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

\( L \): suitable differential operator

- Theory of reproducing kernel Hilbert spaces \((\text{Aronszajn 1950})\)
  \[ \langle f, g \rangle_H = \langle L f, L g \rangle \]

- Interpolation and approximation theory
  - Smoothing splines \((\text{Schoenberg 1964, Kimeldorf-Wahba 1971})\)
  - Thin-plate splines, radial basis functions \((\text{Duchon 1977})\)

- Machine learning
  - Radial basis functions, kernel methods \((\text{Poggio-Girosi 1990})\)
  - Representer theorem(s) \((\text{Schölkopf-Smola 2001})\)

Splines are analog, but intrinsically sparse

\( L\{\cdot\} \): admissible differential operator
\( \delta(\cdot - x_0) \): Dirac impulse shifted by \( x_0 \in \mathbb{R}^d \)

**Definition**

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

\[
L\{s\} = \sum_{k=1}^{K} a_k \delta(\cdot - x_k) = w_\delta : \text{spline’s innovation}
\]

Spline theory: \((\text{Schultz-Varga, 1967})\)
**Spline synthesis: example**

\[ L = D = \frac{d}{dx} \]

Null space: \[ \mathcal{N}_D = \text{span}\{p_1\}, \quad p_1(x) = 1 \]

\[ \rho_D(x) = D^{-1}\{\delta\}(x) = \mathbb{1}_+(x): \text{Heaviside function} \]

\[ w_\delta(x) = \sum_k a_k \delta(x - x_k) \]

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

**Spline synthesis: generalization**

\( L: \) spline admissible operator (LSI)

\[ \rho_L(x) = L^{-1}\{\delta\}(x): \text{Green's function of } L \]

Finite-dimensional null space: \[ \mathcal{N}_L = \text{span}\{p_n\}_{n=1}^{N_0} \]

Spline's innovation:

\[ w_\delta(x) = \sum_k a_k \delta(x - x_k) \]

\[ \Rightarrow s(x) = \sum_k a_k \rho_L(x - x_k) + \sum_{n=1}^{N_0} b_n p_n(x) \]

Requires specification of boundary conditions
RKHS representer theorem for \( L_2 \) regularization

\[
(P2) \quad \text{arg min}_{f \in \mathcal{H}} \left( \sum_{m=1}^{M} |y_m - f(x_m)|^2 + \lambda \|f\|_H^2 \right)
\]

\( r_H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \) is the (unique) reproducing kernel for the Hilbert \( \mathcal{H} \) if
- \( r_H(x_0, \cdot) \in \mathcal{H} \) for all \( x_0 \in \mathbb{R}^d \)
- \( f(x_0) = \langle r_H(x_0, \cdot), f \rangle_H \) for all \( f \in \mathcal{H} \) and \( x_0 \in \mathbb{R}^d \)

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

\[
(P2') \quad \text{arg min}_{f \in \mathcal{H}} \left( F(y, f) + \lambda \|f\|_H^2 \right) \quad \text{(Schölkopf-Smola 2001)}
\]

**Representer theorem for \( L_2 \)-regularization**

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

\[
f(x) = \sum_{m=1}^{M} a_m r_H(x, x_m)
\]

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

L\(_2\) representer theorem for variational splines

**Theoretical difficulty:** \( \|f\|_H^2 \longrightarrow \|L_f\|_{L_2}^2 \) (only a semi-norm !)

\[
(P2) \quad \text{arg min}_{f \in \mathcal{H}_L} \left( \sum_{m=1}^{M} |y_m - f(x_m)|^2 + \lambda \|L_f\|_{L_2(\mathbb{R}^d)}^2 \right)
\]

\( \rho_{L^*L}(x) = (L^*L)^{-1}\{\delta\}(x) \): Green’s function of \( (L^*L) \)

\( (\text{Schoenberg 1964, Kimeldorf-Wahba 1971}) \)

**L\(_2\) representer theorem for variational splines**

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

\[
f(x) = \sum_{m=1}^{M} a_m \rho_{L^*L}(x - x_m) + \sum_{n=1}^{N_0} b_n p_n(x);
\]

i.e., it is a \( (L^*L) \)-spline with knots at the \( \{x_m\} \).

Example: \( L = D^2 \) with \( \rho_{D^4}(x) \propto |x|^3 \) \( \Rightarrow \) \( f(x) \) is a cubic spline
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; Candès & 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 $\ell_1(\mathbb{Z}^d)$

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

- Space of bounded Radon measures on $\mathbb{R}^d$
  \[
  \mathcal{M}(\mathbb{R}^d) = (C_0(\mathbb{R}^d))' = \{ w \in S'(\mathbb{R}^d) : \| w \|_{\mathcal{M}} = \sup_{\varphi \in S(\mathbb{R}^d) : \| \varphi \|_{\infty} = 1} \langle w, \varphi \rangle < \infty \},
  \]
  where \( w : \varphi \mapsto \langle w, \varphi \rangle = \int_{\mathbb{R}^d} \varphi(r)w(r)dr \)

- 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
  - $\delta(\cdot - x_0) \in \mathcal{M}(\mathbb{R}^d)$ with $\| \delta(\cdot - x_0) \|_{\mathcal{M}} = 1$ for any $x_0 \in \mathbb{R}^d$
  - $\| f \|_{\mathcal{M}} = \| f \|_{L_1(\mathbb{R}^d)}$ for all $f \in L_1(\mathbb{R}^d) \Rightarrow L_1(\mathbb{R}^d) \subseteq \mathcal{M}(\mathbb{R}^d)$
Representer theorem for gTV regularization

\[
\begin{align*}
(P1) \quad \arg \min_{f \in \mathcal{M}_L(\mathbb{R}^d)} \left( \sum_{m=1}^M |g_m - \langle h_m, f \rangle|^2 + \lambda \left\| \mathbf{L}f \right\|_{\mathcal{M}} \right)
\end{align*}
\]

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

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

\[
\begin{align*}
(P1') \quad \arg \min_{f \in \mathcal{M}_L(\mathbb{R}^d)} \left( F(\mathbf{y}, \mathbf{\nu}(f)) + \lambda \left\| \mathbf{L}f \right\|_{\mathcal{M}} \right) \quad \text{with} \quad \mathbf{\nu}(f) = (\langle h_1, f \rangle, \ldots, \langle h_M, f \rangle)
\end{align*}
\]

Representer theorem for gTV-regularization

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

\[
f_{\text{spline}}(x) = \sum_{k=1}^{K_{\text{knots}}} a_k \rho_L(x - x_k) + \sum_{n=1}^{N_0} b_n p_n(x)
\]

with \(\rho_L\) such that \(L\{\rho_L\} = \delta\), \(K_{\text{knots}} \leq M - N_0\), and \(\|L f_{\text{spline}}\|_{\mathcal{M}} = \|a\|_{\ell_1}\).


Example: 1D inverse problem with TV(2) regularization

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

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

\[
L = D^2 = \frac{d^2}{dx^2} \quad \rho_{D^2}(x) = (x)_+ : \text{ReLU} \quad \mathcal{N}_{D^2} = \text{span}\{1, x\}
\]

- Generic form of the solution

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

with \(K < M\) and free parameters \(b_1, b_2\) and \((a_k, \tau_k)_{k=1}^K\)
Other spline-admissible operators

- $L = D^n$ (pure derivatives)
  \[ \Rightarrow \text{polynomial splines of degree } (n - 1) \]  
  (Schoenberg 1946)

- $L = D^n + a_{n-1} D^{n-1} + \cdots + a_0 I$ (ordinary differential operator)
  \[ \Rightarrow \text{exponential splines} \]  
  (Dahmen-Micchelli 1987)

- Fractional derivatives: $L = D^\gamma \leftrightarrow \mathcal{F} (j\omega)^\gamma$
  \[ \Rightarrow \text{fractional splines} \]  
  (U.-Blu 2000)

- Fractional Laplacian: $(-\Delta)^{\frac{\gamma}{2}} \leftrightarrow \mathcal{F} \|\omega\|^\gamma$
  \[ \Rightarrow \text{polyharmonic splines} \]  
  (Duchon 1977)

- Elliptical differential operators; e.g., $L = (-\Delta + \alpha I)^\gamma$
  \[ \Rightarrow \text{Sobolev splines} \]  
  (Ward-U. 2014)

**Discretization: compatible with CS paradigm**

\[
s_{\text{sparse}} = \arg \min_{s \in \mathbb{R}^K} \left( \frac{1}{2} \| y - Hs \|^2_2 + \lambda \| u \|_1 \right) \text{ subject to } u = Ls
\]

ADMM algorithm

\[
\mathcal{L}_A(s, u, \alpha) = \frac{1}{2} \| y - Hs \|^2_2 + \lambda \sum_n |[u]_n| + \alpha^T (Ls - u) + \frac{\mu}{2} \| Ls - u \|^2_2
\]

For $k = 0, \ldots, K$

**Linear step**

\[
s^{k+1} = \left( H^T H + \mu L^T L \right)^{-1} (z_0 + z^{k+1})
\]

with \[ z^{k+1} = L^T (\mu u^k - \alpha^k) \]

\[
\alpha^{k+1} = \alpha^k + \mu \left( Ls^{k+1} - u^k \right)
\]

**Proximal step**  = pointwise non-linearity

\[
u^{k+1} = \text{prox}_| \left( Ls^{k+1} + \frac{1}{\mu} \alpha^{k+1}, \frac{\sigma^2}{\mu} \right)
\]
Example: ISMRM reconstruction challenge

$L_2$ regularization (Laplacian) vs. $\ell_1$ / 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 $H^T H$ and $L$ are both convolutions

$$L_A(s, u, \alpha) = \frac{1}{2} \|y - Hs\|_2^2 + \sigma^2 \sum_n |u_n| + \alpha^T (Ls - u) + \frac{\mu}{2} \|Ls - u\|_2^2$$

ADMM algorithm

For $k = 0, \ldots, K$

**Linear step** = Convolutions

$$s^{k+1} = (H^T H + \mu L^T L)^{-1} \left( z_0 + z^{k+1} \right)$$

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

$$\alpha^{k+1} = \alpha^k + \mu \left( Ls^{k+1} - u^k \right)$$

**Proximal step** = pointwise non-linearity

$$u^{k+1} = \text{prox}_{\frac{1}{\mu}} \left( Ls^{k+1} + \frac{1}{\mu} \alpha^{k+1} + \sigma^2 \right)$$

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)
Dose reduction by 7: 143 views

Reconstructed from 1000 views

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

Dose reduction by 20: 50 views

Reconstructed from 1000 views

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

Representer theorem for deep learning

Deep neural networks and splines

- Preferred choice of activation function: ReLU
  - ReLU works nicely with dropout / $\ell_1$-regularization (Glorot ICASSP 2011)
  - Networks with hidden ReLU are easier to train

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

- Linear step: $\mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$
  $f_\ell : x \mapsto f_\ell(x) = W_\ell x + b_\ell$

- Nonlinear step: $\mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$
  $\sigma_\ell : x \mapsto \sigma_\ell(x) = (\sigma_{1,\ell}(x_1), \ldots, \sigma_{N_\ell,\ell}(x_{N_\ell}))$

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

\[
 f_{\text{deep}}(x) = (\sigma_L \circ f_L \circ \sigma_{L-1} \circ \cdots \circ \sigma_2 \circ f_2 \circ \sigma_1 \circ f_1)(x)
\]

Deep neural net with optimized activations

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

- Linear step: $\mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$
  $f_\ell : x \mapsto f_\ell(x) = W_\ell x + b_\ell$

- Nonlinear step: $\mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$
  $\sigma_\ell : x \mapsto \sigma_\ell(x) = (\sigma_{1,\ell}(x_1), \ldots, \sigma_{N_\ell,\ell}(x_{N_\ell}))$

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

\[
 f_{\text{deep}}(x) = (\sigma_L \circ f_L \circ \sigma_{L-1} \circ \cdots \circ \sigma_2 \circ f_2 \circ \sigma_1 \circ f_1)(x)
\]
Theorem (TV(2)-optimality of deep spline networks)

- Neural network \( f : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L} \) with deep structure \((N_0, N_1, \ldots, N_L)\)
  \( x \mapsto f(x) = (\sigma_L \circ \ell_L \circ \sigma_{L-1} \circ \cdots \circ \ell_2 \circ \sigma_1 \circ \ell_1)(x) \)
- Normalized linear transformations \( \ell_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_{\ell}}, x \mapsto U_\ell x \) with weights \( U_\ell = [u_{1,\ell} \cdots u_{N_{\ell}}]^{T} \in \mathbb{R}^{N_{\ell} \times N_{\ell-1}} \) such that \( \|u_{\ell,1}\| = 1 \)
- Free-form activations \( \sigma_\ell = (\sigma_{1,\ell}, \ldots, \sigma_{N_{\ell},\ell}) : \mathbb{R}^{N_{\ell}} \to \mathbb{R}^{N_{\ell}} \) with \( \sigma_{1,\ell}, \ldots, \sigma_{N_{\ell},\ell} \in \text{BV}(2)(\mathbb{R}) \)

Given a series of \( M \) data points \( y_m \approx f(x_m) \), we then define the training problem

\[
\text{arg min}_{(U_\ell), (\sigma_{n,\ell} \in \text{BV}(2)(\mathbb{R}))} \left( \sum_{m=1}^{M} E(y_m, f(x_m)) + \mu \sum_{\ell=1}^{N} R_\ell(U_\ell) + \lambda \sum_{\ell=1}^{L} \sum_{n=1}^{N_\ell} \text{TV}(2)(\sigma_{n,\ell}) \right)
\]

where

- \( E : \mathbb{R}^{N_L} \times \mathbb{R}^{N_L} \to \mathbb{R}^+ \): 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 \( K_{n,\ell} \leq M - 2 \), \( \tau_{1,n,\ell}, \ldots, \tau_{K_{n,\ell},n,\ell} \in \mathbb{R} \), and \( b_{1,n,\ell}, b_{2,n,\ell}, a_{1,n,\ell}, \ldots, a_{K_{n,\ell},n,\ell} \in \mathbb{R} \).

Outcome of representer theorem

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

- its number \( 0 \leq K = K_{n,\ell} \) 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 \( b_{n,\ell} = (b_{1,n,\ell}, b_{2,n,\ell}) \in \mathbb{R}^2 \), \( a_{n,\ell} = (a_{1,n,\ell}, \ldots, a_{K,n,\ell}) \in \mathbb{R}^K \).

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

- Link with \( \ell_1 \) minimization techniques

\[
\text{TV}(2)(\sigma_{n,\ell}) = \sum_{k=1}^{K_{n,\ell}} |a_{k,n,\ell}| = \|a_{n,\ell}\|_1
\]
Comparison of linear interpolators

Deep spline networks: Discussion

- Global optimality achieved with spline activations
- State-of-the-art ReLU networks \((K_{n,\ell} = 1, b_{n,\ell} = 0)\)

  - No need to normalize:
    \[
    (w^T_{n,\ell} x - z_{n,\ell})_+ = (a_{n,\ell} u^T_{n,\ell} x - z_{n,\ell})_+ = a_{n,\ell} (u^T_{n,\ell} x - \tau_{n,\ell})_+
    \]

- Key features
  - Produces a global mapping \(x \mapsto f(x)\) that is continuous and piecewise-linear
  - Direct control of complexity (number of knots): adjustment of \(\lambda\)
  - Ability to suppress unnecessary layers

- Backward compatibility
  - Linear regression: \(\lambda \to \infty \Rightarrow K_{n,\ell} = 0\)
  - Compressed sensing / \(\ell_1\) 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
  - Functional model: class of signals + physics
  - Smoothing-splines: minimum “spline” energy
    \[ (L^*L)\{s_{\text{smooth}}\} = \sum_{m=1}^{M} a_m h_m \]
  - L-splines = signals with “sparsest” innovation
    \[ L\{s_{\text{sparse}}\} = \sum_{k=1}^{K} a_k \delta(\cdot - x_k) \]

- 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**

- **Theory of sparse stochastic processes**
  - For splines: see chapter 6

- **Deep neural networks**