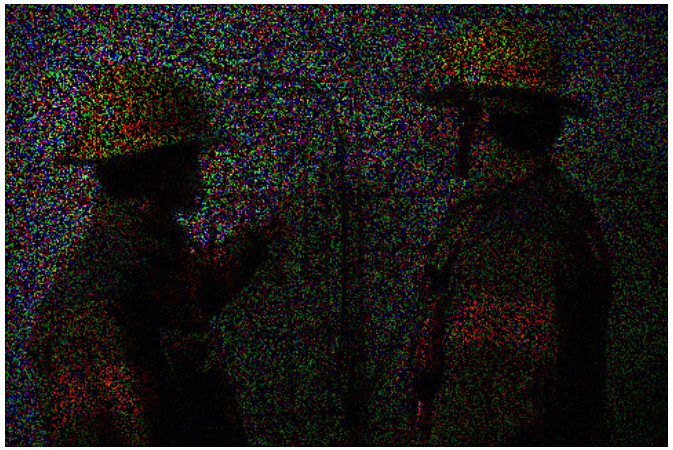
Inverse problems and optimization

Nelly PUSTELNIK
CNRS, Laboratoire de Physique de l'ENS Lyon
Laurent CONDAT
CNRS, Gipsa-lab

Course 4: Sparsity
January, 19th 2017

Motivation: image restoration

$\overline{x} \in \mathbb{R}^N$



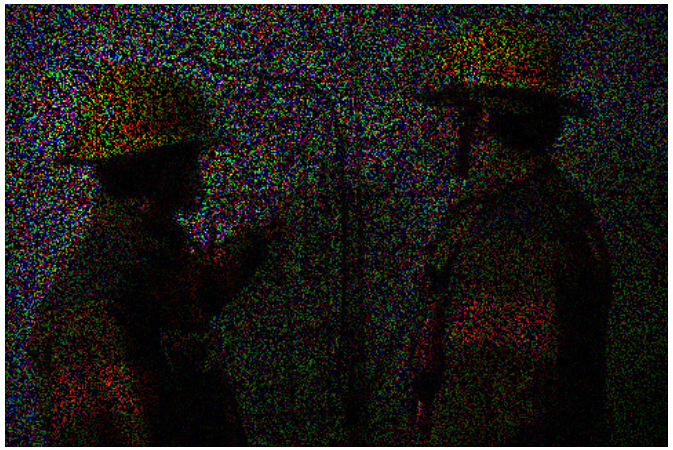
$z \in \mathbb{R}^M$

Objective recover $\overline{x}$ from the observations $z = \mathcal{P}_\alpha(H\overline{x})$

- $\overline{x}$: original image
- H : linear operator from $\mathbb{R}^N$ to $\mathbb{R}^M$
- $\mathcal{P}_\alpha$: effect of noise where $\alpha > 0$ is the scaling parameter
- z : degraded image of size M

Motivation: image restoration

$\overline{x} \in \mathbb{R}^N$



$z \in \mathbb{R}^M$

Objective recover $\overline{x}$ from the observations $z = \mathcal{P}_\alpha(H\overline{x})$

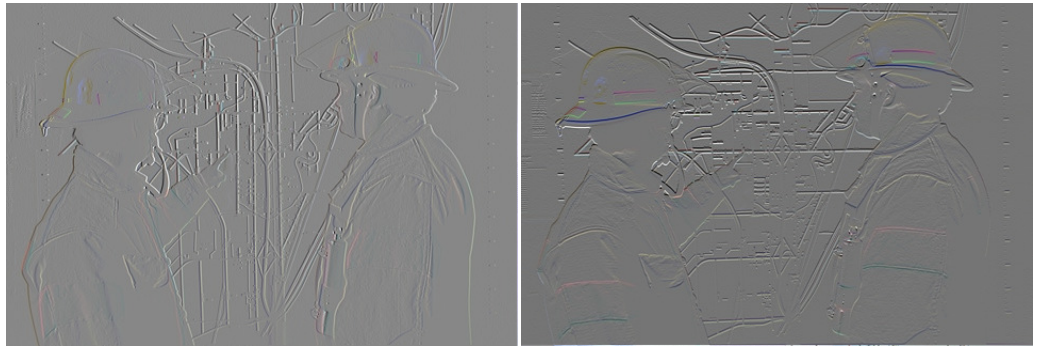
- $\overline{x}$: original image $\Rightarrow$ sparse after some appropriate transform
- H : linear operator from $\mathbb{R}^N$ to $\mathbb{R}^M$
- $\mathcal{P}_\alpha$: effect of noise where $\alpha > 0$ is the scaling parameter
- z : degraded image of size M

Sparse transforms: finite difference

$x \in \mathbb{R}^N$

F

F^*



$\theta \in \mathbb{R}^K$

Sparsity coefficients $\theta = Fx$

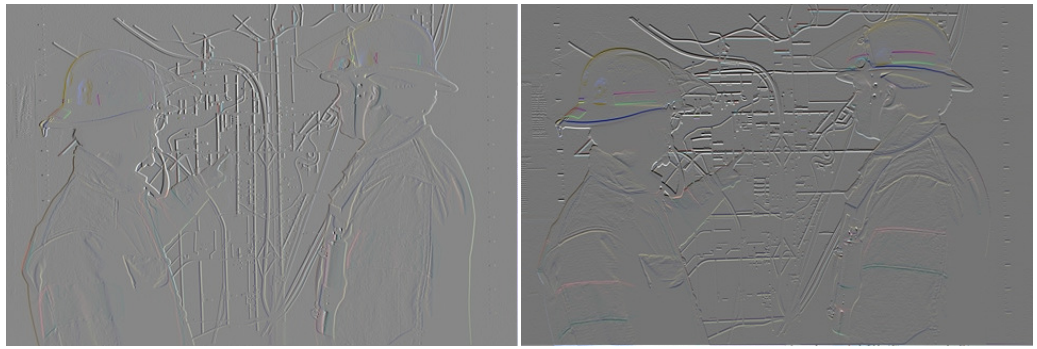
- θ : coefficients of size $K \geq N$
- $F \in \mathbb{R}^{K \times N}$: matrix associated with the analysis operator
- $F^* \in \mathbb{R}^{N \times K}$: matrix associated with the synthesis operator.

Sparse transforms: finite difference

$x \in \mathbb{R}^N$

F

F^*



$\theta \in \mathbb{R}^K$

- Horizontal and vertical difference filters: $f_1 = [1 - 1]$ and $f_2 = f_1\top$,
- $K = 2N$,
- Link between (f_1, f_2) and (F_1, F_2): cf. slides10-12 course 1,
- Resulting $F = [F_1^\top, F_2^\top]^\top$ and $F^* = F^\top$,
- $F^*F \neq \mathrm{Id}$.

Sparse transforms: wavelet transform

$x \in \mathbb{R}^N$

F

F^*

$\theta \in \mathbb{R}^N$

Sparsity coefficients $\theta = Fx$

- θ : coefficients of size N
- $F \in \mathbb{R}^{N \times N}$: matrix associated with the analysis operator
- $F^* \in \mathbb{R}^{N \times N}$: matrix associated with the synthesis operator.

$\Rightarrow$ Orthonormal transform: $F^*F = \mathrm{Id}$ for $\mu > 0$.

Sparse transforms: wavelet transform

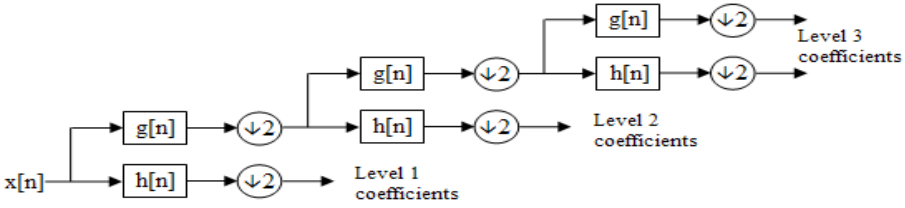
$$x \in \mathbb{R}^N$$

$$F$$

$$F^*$$

$$\theta \in \mathbb{R}^K$$

- F (resp. F^*) : Concatenation of matrices associated to filtering and decimation (resp. upsampling) operations.



Sparse transforms: frame transform

$x \in \mathbb{R}^N$

F

F^*

$\theta \in \mathbb{R}^K$

Sparsity coefficients $\theta = Fx$

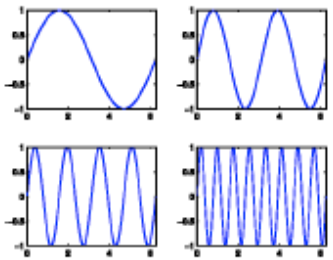
- θ : coefficients of size $K \geq N$
- $F \in \mathbb{R}^{K \times N}$: matrix associated with the analysis operator
- $F^* \in \mathbb{R}^{N \times K}$: matrix associated with the synthesis operator.

$\Rightarrow$ Tight frame condition: $F^*F = \mu \mathrm{Id}$ for $\mu > 0$.

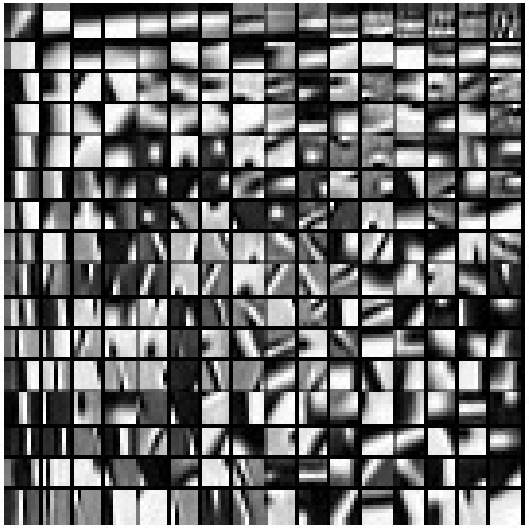
Sparse transforms: dictionary

- Dictionary: set of elementary signals

$$D = [d_1, \ldots, d_K] \in \mathbb{R}^{N \times K}$$



- D is "adapted" to x if it can represent it with a few elements, i.e., there exists a sparse vector θ in $\mathbb{R}^K$ such that $x \sim D\theta$.

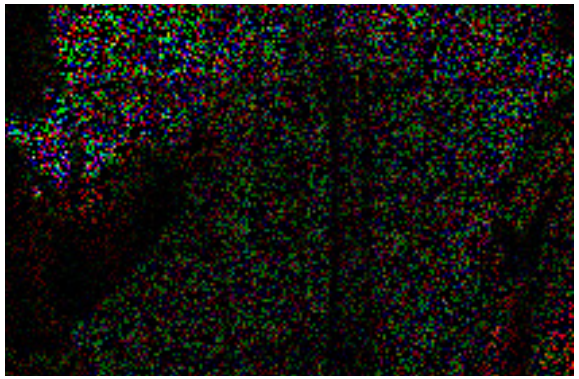


- θ: sparse code.

Sparse transforms

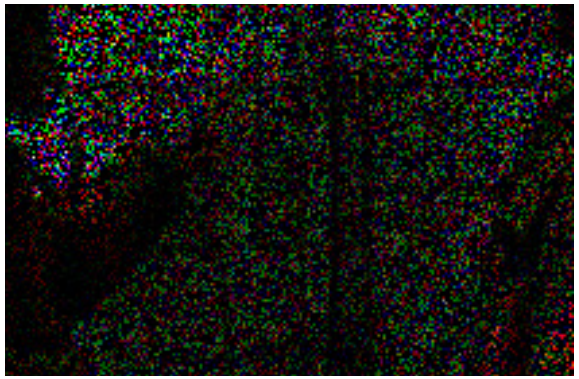
Quadratic regularization - Wiener filtering [Wiener,1949]

$$\widehat{y} = \underset{y \in \mathbb{R}^N}{\operatorname{argmin}} \|Hx - z\|_2^2 + \lambda \|Fx\|_2^2 \qquad \lambda > 0$$



Use of frames - Soft-thresholding [Haar,1910] [Mallat,2009]

$$\widehat{y} = \underset{y \in \mathbb{R}^N}{\operatorname{argmin}} \|Hx - z\|_2^2 + \lambda \|Fx\|_1 \qquad \lambda > 0$$



- The squared ℓ_2-norm induces "smoothness" while ℓ_1-norm induces sparsity. [Chen et al., 1999, Tibshirani, 1996]

Sparse transforms

Quadratic regularization - Wiener filtering [Wiener,1949]

$$\widehat{y} = \underset{y \in \mathbb{R}^N}{\operatorname{argmin}} \|Hx - z\|_2^2 + \lambda \|Fx\|_2^2 \qquad \lambda > 0$$

Use of frames - Soft-thresholding [Haar,1910] [Mallat,2009]

$$\widehat{y} = \underset{y \in \mathbb{R}^N}{\operatorname{argmin}} \|Hx - z\|_2^2 + \lambda \|Fx\|_1 \qquad \lambda > 0$$

- The squared ℓ_2-norm induces "smoothness" while ℓ_1-norm induces sparsity. [Chen et al., 1999, Tibshirani, 1996]

Sparse transforms: analysis/synthesis

Analysis formulation

$$\widehat{x} \in \underset{x \in \mathbb{R}^N}{\mathrm{Argmin}} \|Hx - z\|_2^2 + \lambda \|Fx\|_1 \qquad \lambda > 0$$

Synthesis formulation

$$\widehat{x} = F^* \widehat{\theta} \quad \text{with} \quad \widehat{\theta} \in \underset{\theta \in \mathbb{R}^K}{\mathrm{Argmin}} \|HF^*\theta - z\|_2^2 + \lambda \|\theta\|_1 \qquad \lambda > 0$$

- Sparse coding with $F^* = D$.
- Equivalence for some specific conditions over F such orthogonality. [Pustelnik et al. 2016]

Variational approach: Bayesian formulation

- $u = Hx = (u^{(i)})_{1 \leq i \leq N}$: realization of a random vector U.
- z: realization of a random vector Z.
- $\theta = Fx = (\theta^{(k)})_{1 \leq k \leq K}$: realization of a random vector $\overline{\Theta} = (\Theta^{(k)})_{1 \leq k \leq K}$ having independent components.

MAP estimator (Maximum A Posteriori)

$$\max_{x} \ \mathrm{P}(U = Hx \mid Z = z)$$

$$\max_{\theta} \ \mathrm{P}(Z = z \mid U = HF^*\theta) \cdot \mathrm{P}(\Theta = \theta)$$

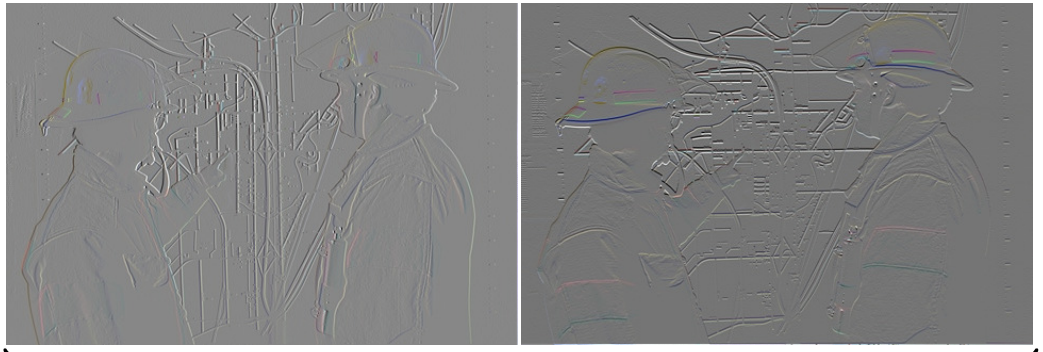
$$\min_{\theta} \ \underbrace{-\ln \mathrm{P}(Z = z \mid U = HF^*\theta)}_{\text{Data fidelity}} \underbrace{- \sum_{k=1}^{K} \ln \mathrm{p}_{\Theta^{(k)}}(\theta^{(k)})}_{\text{A priori}}$$

Variational approach: Bayesian formulation

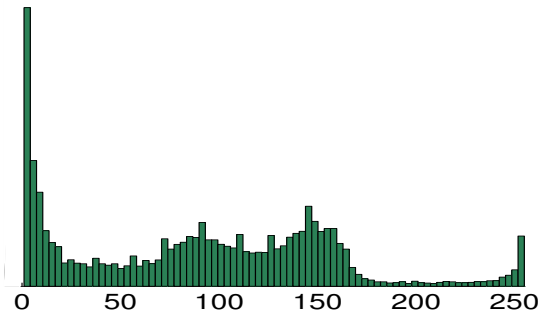
$$x \in \mathbb{R}^N$$

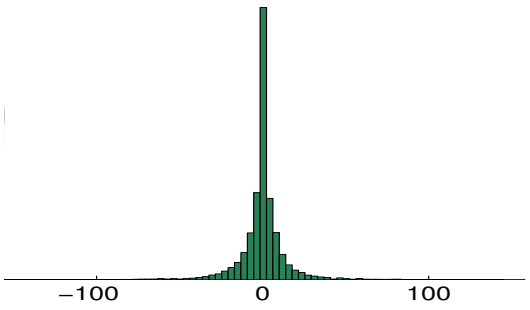
F

F^*



$$\theta \in \mathbb{R}^K$$





Variational approach: Bayesian formulation

MAP estimator (Maximum A Posteriori)

$$\min_x \quad -\underbrace{\ln \mathrm{P}(Z = z \mid U = HF^*\theta)}_{\text{Data fidelity}} - \underbrace{\sum_{k=1}^{K} \ln \mathrm{p}_{\Theta^{(k)}}(\theta^{(k)})}_{\text{A priori}}$$

where

$$\mathrm{P}(Z = z \mid U = AF^*\theta) = \frac{1}{(2\pi\alpha)^{M/2}} \exp\left\{ -\frac{\|HF^*\theta - z\|_2^2}{2\alpha} \right\}$$

and

$$p_{\Theta^{(k)}}(\theta^{(k)}) = \frac{1}{C_k} \exp\{-\lambda_k |\theta^{(k)}|\}$$

$$\min_\theta \quad \frac{1}{2\alpha}\|HF^*\theta - z\|_2^2 + \sum_{k=1}^{K} \lambda_k |\theta^{(k)}|$$

Variational approach: Equivalent formulations

- Geophysics: [Claerbout and Muir, 1973, Taylor et al., 1979],
- Statistics: Lasso [Tibshirani, 1996]
- Signal processing: Basis pursuit [Chen et al., 1999]

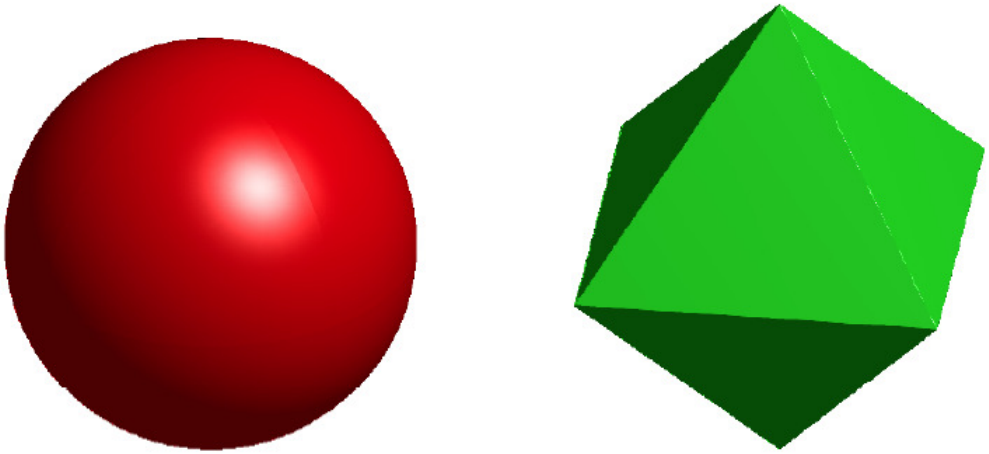
Equivalent formulations:

$$\min_{\theta \in \mathbb{R}^K} \quad \frac{1}{2\alpha} \|HF^*\theta - z\|_2^2 + \lambda \|\theta\|_1$$

$$\min_{\theta \in \mathbb{R}^K} \quad \frac{1}{2\alpha} \|HF^*\theta - z\|_2^2 \quad \text{s.t.} \quad \|\theta\|_1 \leq \mu$$

$$\min_{\theta \in \mathbb{R}^K} \quad \|\theta\|_1 \quad \text{s.t.} \quad \|HF^*\theta - z\|_2^2 \leq \varepsilon$$

Why does the ℓ_1-norm induce sparsity ?



(figures extracted from J. Mairal course)

Why does the ℓ_1-norm induce sparsity ?

$$\min_{\theta \in \mathbb{R}} \quad \frac{1}{2}(\theta - z)^2 + \lambda|\theta|$$

- Piecewise quadratic function with a non-differentiability at zero.
- Fermat rule for non-differentiable function:

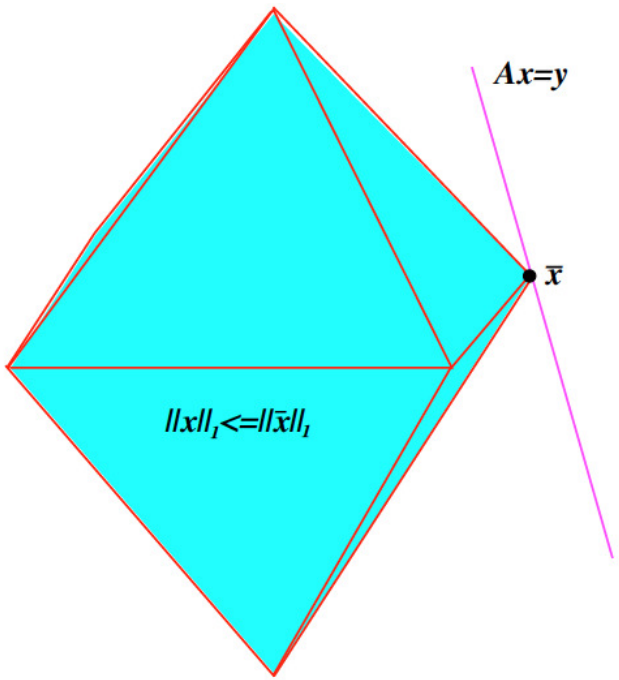
$$0 \in \widehat{\theta} - z + \lambda u \quad \text{with} \quad u \in \partial|\widehat{\theta}|$$

- This lead to soft-thresholding

$$\widehat{\theta} = \mathrm{sign}(z)\max(0, |z| - \lambda)$$

Why does the ℓ_1-norm induce sparsity ?

- Resolution of $\min_{x \in \mathbb{R}^K} \quad \|x\|_1 \quad \text{s.t.} \quad Ax = y$



Why does the ℓ_1-norm induce sparsity ?

- ℓ_2^2 versus ℓ_1

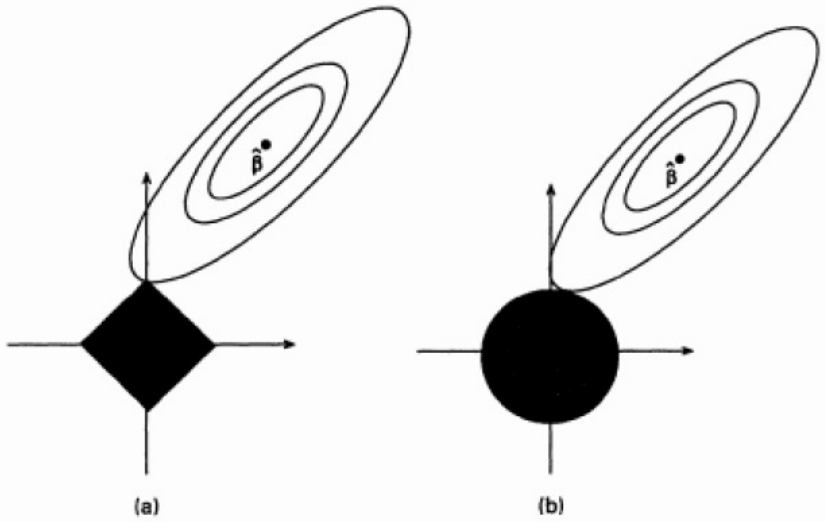


Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

Non-convex functions

- ℓ_q-norms with $q \in]0,1[$ [Frank and Friedman, 1993]

$$\psi(\theta) = \sum_{k=1}^{K} |\theta_k|^q$$

- Log penalty [Candès et al. 2008]

$$\psi(\theta) = \sum_{k=1}^{K} \log(|\theta_k| + \varepsilon)$$

- Several others [Nikolova, 2007]

- Non-convex penalties leading to convex criterion [Selesnick et al. 2015]

Compressed sensing

Sparsity assumption

$$\theta_0 \in \arg\min_{\theta \in \mathbb{R}^K} \frac{1}{2\alpha} \|D\theta - z\|_2^2 + \lambda \|\theta\|_0$$

where $(\forall \theta = (\theta_k)_{1 \leq k \leq K} \in \mathbb{R}^K)$ $\|\theta\|_0 \equiv \#\{k : \theta_k \neq 0\}$

Convex relaxation

$$\widehat{\theta}_0 \in \arg\min_{\theta \in \mathbb{R}^K} \frac{1}{2\alpha} \|D\theta - z\|_2^2 + \lambda \|\theta\|_1$$

where $(\forall \theta = (\theta_k)_{1 \leq k \leq K} \in \mathbb{R}^K)$ $\|\theta\|_1 = \sum_{k=1}^{N} |\theta_k|$

$\Rightarrow$ Compressed sensing [Donoho,2004][Candès et al. 2006]

Compressed sensing

- **Restricted Isometry Property** (RIP) [Candès, 2008]:

$$(1-\delta_{2s}^{\min})\|\theta\|^2 \leq \|D\theta\|^2 \leq (1+\delta_{2s}^{\max})\|\theta\|^2$$

where

$$\begin{cases} \delta_s^{\min} = 1 - \lambda^{\min}(D_I^* D_I) \\ \delta_s^{\max} = \lambda^{\max}(D_I^* D_I) - 1 \end{cases} \quad \text{and} \quad D_I = (d_i)_{i\in I} \text{ with } |I| = s.$$

Theorem [Candès, 2008]

Assume that $\delta_{2s} = \min(\delta_{2s}^{\min}, \delta_{2s}^{\max}) < \sqrt{2}-1$. Then the solution $\hat{x}$ to the ℓ_1-minimization obeys

$$\|\hat{\theta}-\theta_0\|_1 \leq C_0\|\theta_0-\theta_s\|_1 \quad \text{and} \quad \|\hat{\theta}-\theta_0\|_2 \leq C_0 s^{-1/2}\|\theta_0-\theta_s\|_1$$

for some constant C_0 and where θ_s denotes the s highest non-zero component of θ_0. In particular, if θ is s-sparse, the recovery is exact.

Compressed sensing

- **Restricted Isometry Property** (RIP) [Candès, 2008]:
  - Extended proposition to insure robustness to noise.
  - Sufficient condition (not necessary).
  - NP-hard to compute δ_{2s}.
  - $\delta_{2s} < \sqrt{2} - 1$: restrictive condition even for random matrices [Dossal et al., 2010].

Compressed sensing

- **Sub-differential of ℓ_1 norm** [Fuchs, 2004][Tropp, 2006]

Theorem [Fuchs, 2004]

The solution θ_0 of $D\theta = z$ can be recovered from the unique optimum point $\widehat{\theta}$ of ℓ_1-minimization if

$$1)\ |d_j^\top d_0| < 1, \quad \forall d_j \notin \overline{D}_0 \ \text{ with } \ d_0 = \overline{D}_0^{+\top} \mathrm{sign}(\overline{\theta}_0),$$

$$2)\ h \in]0, h_m[\ \text{ the domain in which}$$

$$\mathrm{sign}\{\overline{\theta}_0 - h(\overline{D}_0^\top \overline{D}_0)^{-1} \mathrm{sign}(\overline{\theta}_0)\} = \mathrm{sign}(\overline{\theta}_0),$$

where

- $\overline{\theta}_0$ is built upon the nonzero component of x_0,
- $\overline{D}_0$ a full rank matrix such that $Dx_0 = \overline{D}_0 \overline{x}_0$,
- $\overline{D}_0^+ = (\overline{D}_0^\top \overline{D}_0)^{-1} \overline{D}_0^\top$ denotes the pseudo-inverse of $\overline{D}$.

Compressed sensing

- **Sub-differential of ℓ_1 norm** [Fuchs, 2004][Tropp, 2006]
  - Extended proposition to ensure robustness to noise.
  - Sufficient condition (not necessary).
  - NP-hard to check the condition
    $|d_j^\top d_0| < 1, \quad \forall d_j \notin \overline{D}_0$ with $d_0 = \overline{D}_0^{+\top} \mathrm{sign}(\overline{\theta}_0)$ for all vector s-sparse.

Compressed sensing

- **Coherence** [Fuchs, 2004]

Theorem [Fuchs, 2004]

The solution θ_0 of $D\theta = z$ can be recovered from the unique optimum point $\widehat{\theta}$ of ℓ_1-minimization if

$$\|\theta_0\|_0 < \frac{1}{2}\left(1 + \frac{1}{M}\right) \qquad \text{where} \qquad M = \max_{1 \le i \ne j \le m} |d_i^\top d_j|.$$

Compressed sensing

- **Coherence** [Fuchs, 2004]
  - Extended proposition to ensure robustness to noise.
  - Sub-differential of ℓ_1 norm condition $\Rightarrow$ Coherence condition.
  - Sufficient condition (not necessary).
  - Easy to compute $\frac{1}{2}\left(1+\frac{1}{M}\right)$ and thus to know the allowed sparsity degree.

Compressed sensing

- **Polytope theory** [Donoho,2004] : NSC
- A large panel of the literature provides results based on greedy algorithms (MP, OMP, OLS) solving rather that l1-minimization. MP [Mallat 93], projection pursuit [Friedman 81], [Huber 85], pure greedy [Temlyakov 08], OMP : [Pati 93], [Zhang 93], [Davis 94], orthogonal greedy algorithm [Temlyakov 08], OLS : [Chen 89], forward selection [Miller 02], greedy algorithm [Natarajan 95], order recursive matching pursuit [Cotter 99], optimized orthogonal matching pursuit [Reibollo-Neira 02], pure orthogonal matching pursuit [Foucart 11].
- **Spectral sparsity** [Fazel et al., 2001] [Srebro et al., 2005] : ℓ_1-ℓ_0 formulation for matrices. Minimization of the rank or its relaxation involving nuclear norm

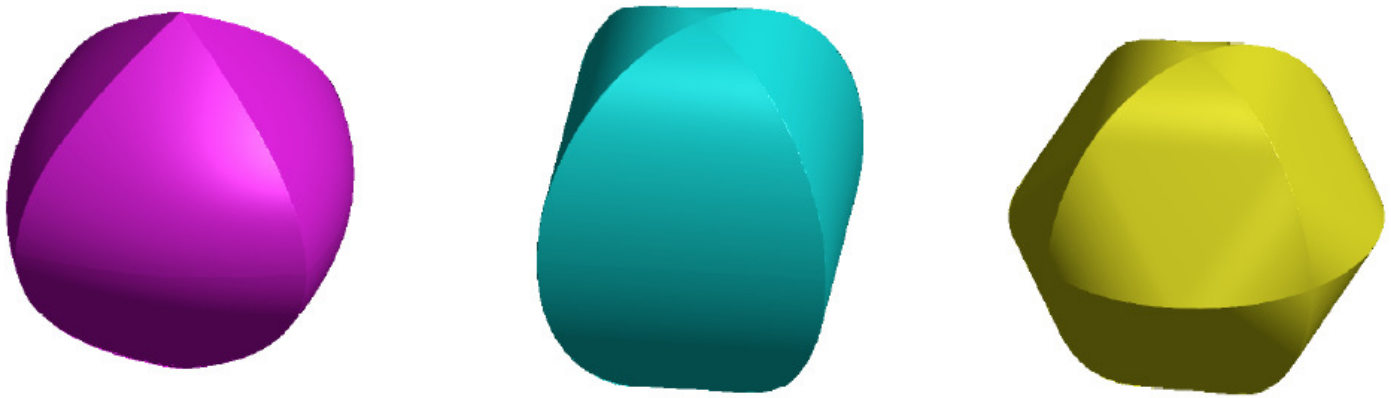
Compressed sensing

- **Difference between random and deterministic matrices:**

  - In a random context, the eigenvalue distribution can be controlled and it results an explicit relation between s, K, and N.

  - In a deterministic context and without prior onto the support or the sign, the validation of the previous properties (RIP, Coherence,...) requires a high computational cost.

Structured sparsity



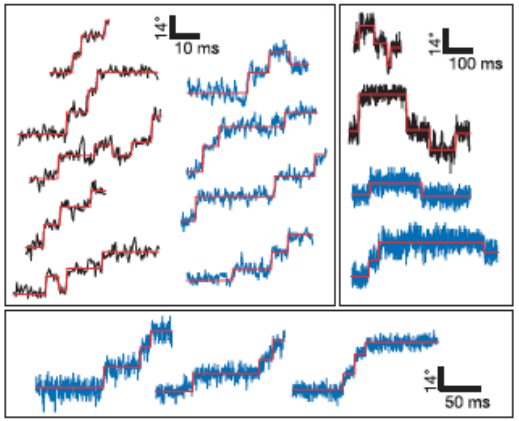
(figures extracted from J. Mairal course)

1D-total variation

Anisotropic total variation [Rudin et al. 1992]

$$\widehat{x} = \arg\min_{x \in \mathbb{R}^N} \|x - z\|_2^2 + \lambda \sum_{n=1}^{N-1} |x_{n+1} - x_n|$$

- Statistics: fused Lasso
  [Tibshirani et al 2005]
- ℓ_0-formulation: Potts model
  [Geman, Geman 1984]
  [Yao 1984]
  [Mumford, Shah 1989]
  [Winkler, Liebscher 2002]



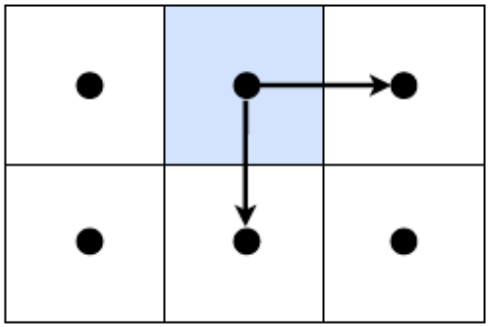
(cf. [Sowa et al. 2005])

2D-total variation

Anisotropic total variation [Rudin et al. 1992]

$$\widehat{x} = \arg\min_{x \in \mathbb{R}^N} \|x - z\|_2^2 + \lambda \sum_{n_1=1}^{N-1} \sum_{n_2=1}^{N-1} \left(|x_{n_1+1,n_2} - x_{n_1,n_2}|^2 + |x_{n_1,n_2+1} - x_{n_1,n_2}|^2 \right)$$

- Horizontal and vertical difference filters:
  $f_1 = [1 \; -1]$ and $f_2 = f_1 \top$
- Link between (f_1, f_2) and (F_1, F_2):
  cf. slides 10-12 course 1
- Sparse transform:
  $F = [F_1^\top, F_2^\top]^\top \in \mathbb{R}^{2N \times N}$
- Regularization:
  $\psi(x) = \|Fx\|_1 = \|F_1 x\|_1 + \|F_2 x\|_1$

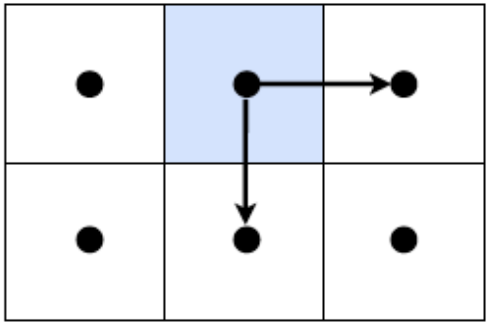


2D-total variation

Isotropic total variation [Rudin et al. 1992]

$$\widehat{x} = \arg\min_{x \in \mathbb{R}^N} \|x - z\|_2^2 + \lambda \sum_{n_1=1}^{N-1} \sum_{n_2=1}^{N-1} \sqrt{|x_{n_1+1,n_2} - x_{n_1,n_2}|^2 + |x_{n_1,n_2+1} - x_{n_1,n_2}|^2}$$

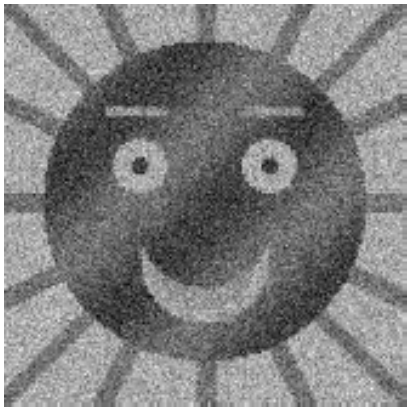
- Horizontal and vertical difference filters: $f_1 = [1 \; -1]$ and $f_2 = f_1\top$
- Link between (f_1, f_2) and (F_1, F_2): cf. slides 10-12 course 1
- Sparse transform: $F = [F_1^\top, F_2^\top]^\top \in \mathbb{R}^{2N \times N}$
- Regularization: $\psi(x) = \|Fx\|_{1,2}$



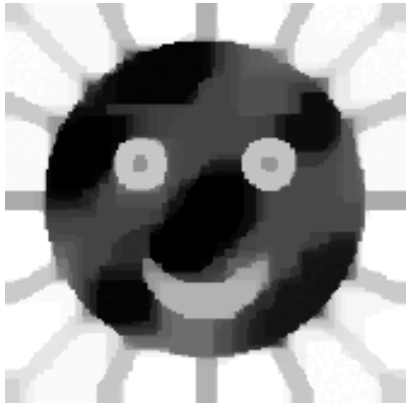
2D-total variation

Original x

Noisy z



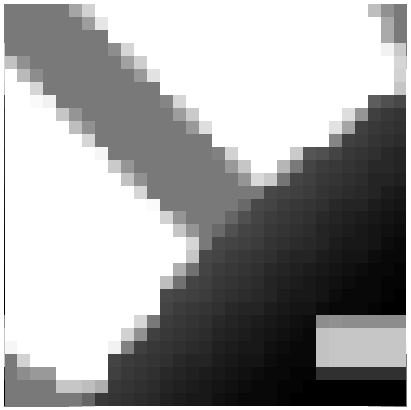
Anisotropic TV



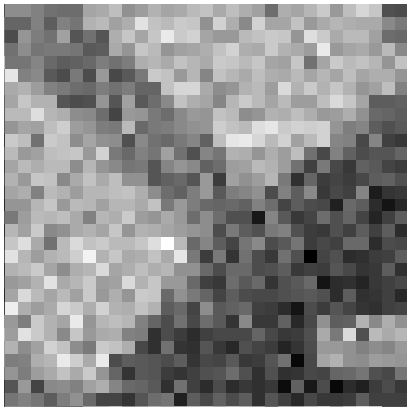
Isotropic TV

2D-total variation

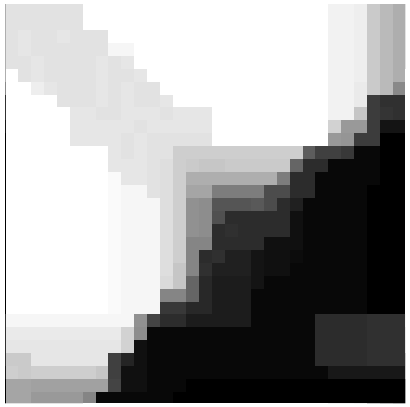
Original x



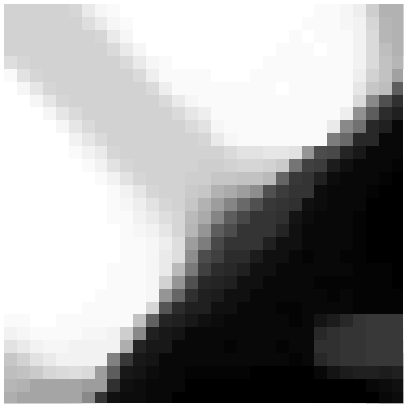
Noisy z



Anisotropic TV



Isotropic TV

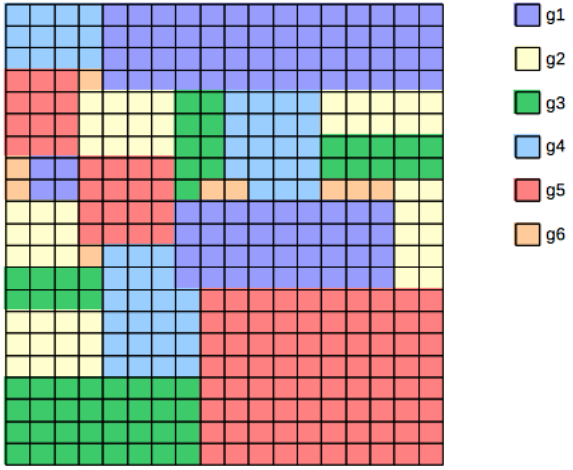


Mixed-norm

Group-lasso [Turlach et al., 2005] [Yuan and Lin, 2006] [Zhao et al., 2009] [Grandvalet and Canu, 1999] [Bakin, 1999]

$$\widehat{x} = \underset{\theta \in \mathbb{R}^K}{\arg\min} \, \|D\theta - z\|_2^2 + \lambda \sum_{g \in \mathcal{G}} \|\theta_g\|_q$$

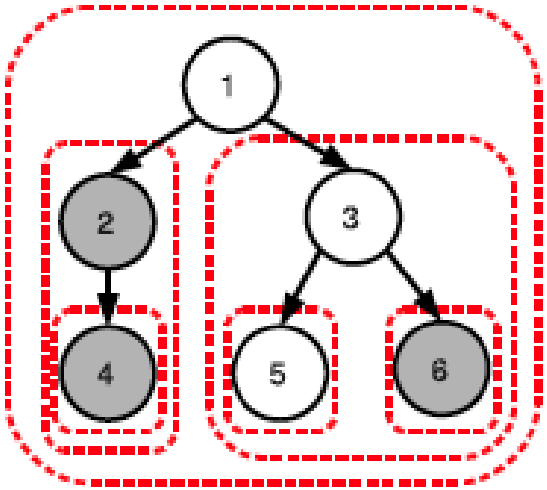
- $\mathcal{G}$ is a partition of $\{1, \ldots, K\}$
- $q = \{2, +\infty\}$
- Can be interpreted as the ℓ_1-norm of $(\|\theta_g\|_q)_{g \in \mathcal{G}}$.
  $\rightarrow$ Non-overlapping groups



Structured sparsity

- Tree-structure [Zhao et al. 2009]
- Select a union of groups [Jacob et al. 2009]
- Zero-pattern in a union of groups [Jenatton et al, 2011]

  $\rightarrow$ Overlapping groups
- ...



Some take-home messages

- Sparsity is not always good. If possible, try ℓ_2^2 before trying ℓ_1.
- The dictionaries used in practice rarely satisfy the assumptions ensuring sparse recovery.
- Dictionary learning consists of estimating D and θ simultaneously.
- There are numerous ways of designing sparse regularization functions adapted to a particular problem. Choosing the best one is not easy and requires some domain knowledge.
- Solving criterion involving ℓ_1-norm requires specific algorithmic procedures described in next courses.