A hand-waving introduction to sparsity for compressed tomography reconstruction

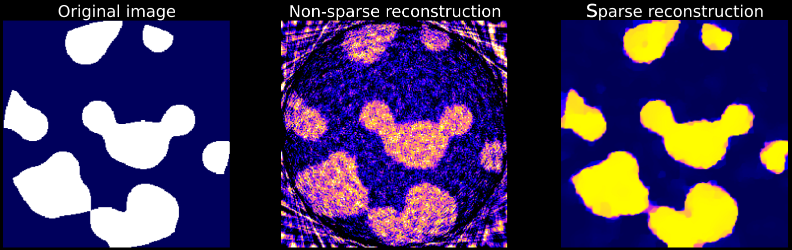
Gaël Varoquaux and **Emmanuelle Gouillart**

PARIETAL

SAINT-GOBAIN

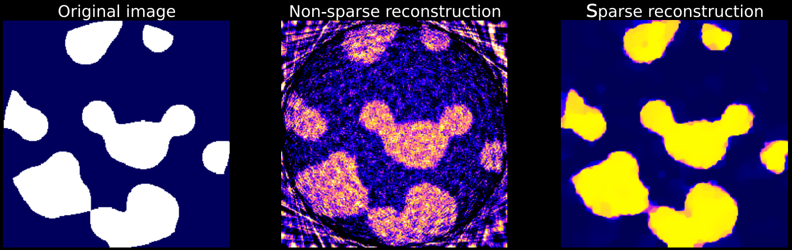
Inria

cnrs



Dense slides. For future reference: `http://www.slideshare.net/GaelVaroquaux`

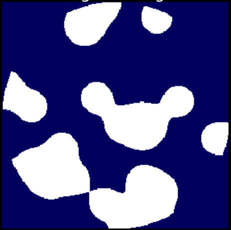
1. Sparsity for inverse problems
2. Mathematical formulation
3. Choice of a sparse representation
4. Optimization algorithms



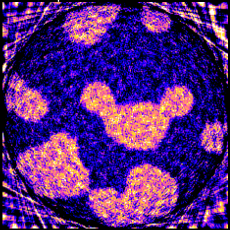
1 Sparsity for inverse problems

- Problem setting
- Intuitions

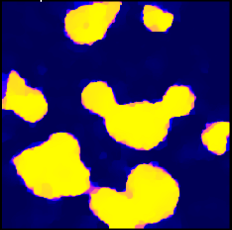
Original image



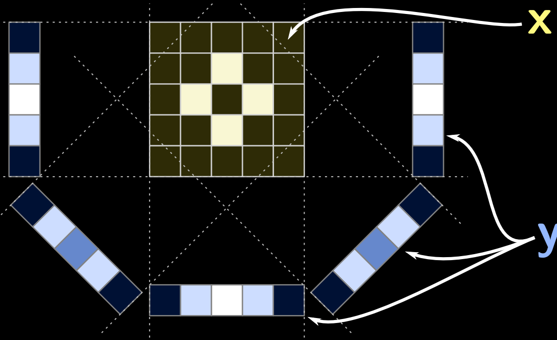
Non-sparse reconstruction



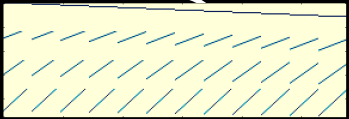
Sparse reconstruction



1 Tomography reconstruction: a linear problem



$$\mathbf{y} = \mathbf{A}\,\mathbf{x}$$



$$\mathbf{y} \in \mathbb{R}^n, \quad \mathbf{A} \in \mathbb{R}^{n \times p}, \quad \mathbf{x} \in \mathbb{R}^p$$

$n \propto$ number of projections

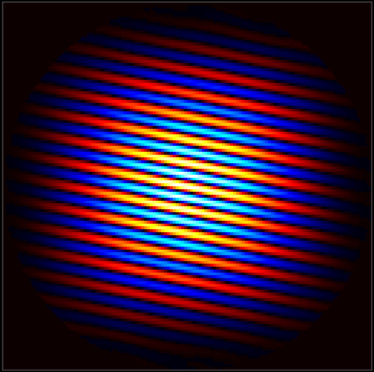
p: number of pixels in reconstructed image

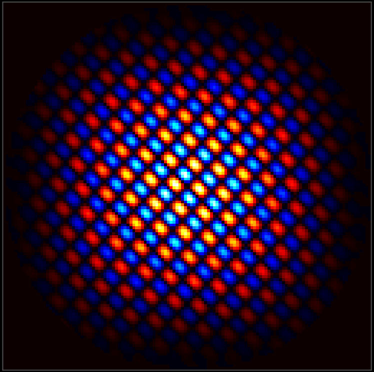
We want to find x knowing A and y

1 Small n: an ill-posed linear problem

$$\mathbf{y} = \mathbf{A}\,\mathbf{x} \quad \text{admits multiple solutions}$$

- The sensing operator **A** has a large null space: images that give null projections
- In particular it is blind to high spatial frequencies:

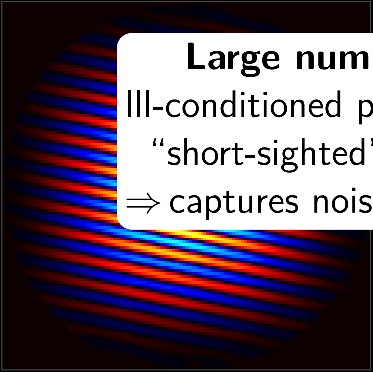


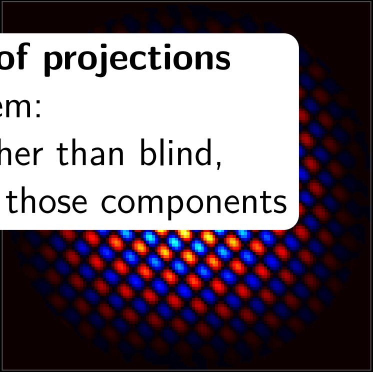


1 Small n: an ill-posed linear problem

$$\mathbf{y} = \mathbf{A}\,\mathbf{x}$$ admits multiple solutions

- The sensing operator $\mathbf{A}$ has a large null space: images that give null projections
- In particular it is blind to high spatial frequencies:





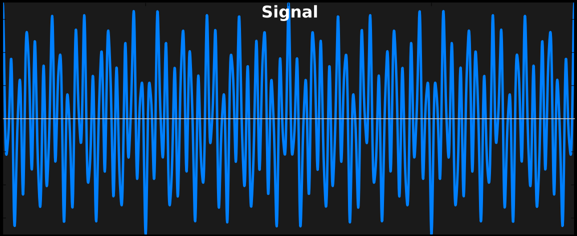
Large number of projections

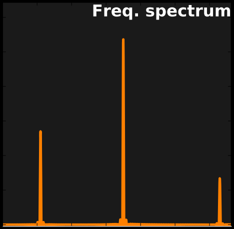
Ill-conditioned problem:

"short-sighted" rather than blind,

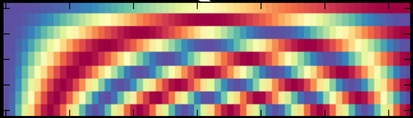
$\Rightarrow$ captures noise on those components

Recovering the frequency spectrum



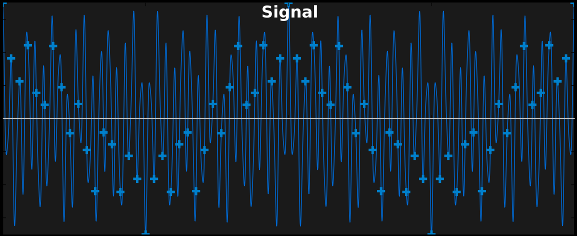


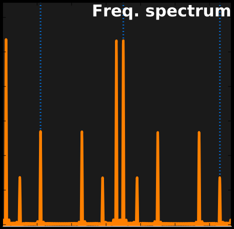
$$\text{signal} = \mathbf{A} \cdot \text{frequencies}$$



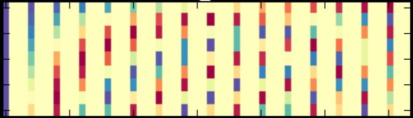
...

Sub-sampling





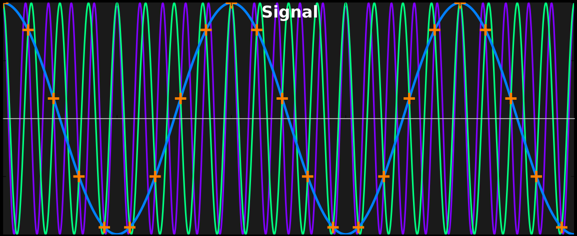
$$\text{signal} = \mathbf{A} \cdot \text{frequencies}$$

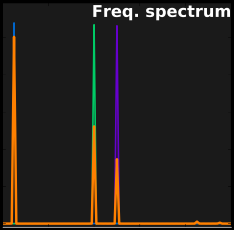


...

- Recovery problem becomes ill-posed

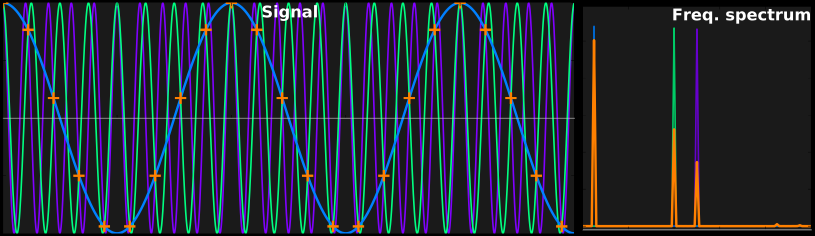
Problem: aliasing



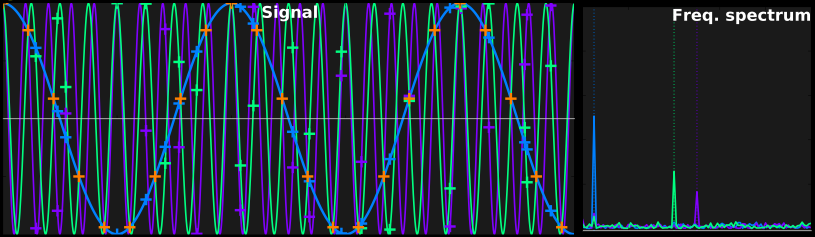


- Information in the null-space of **A** is lost

Problem: aliasing

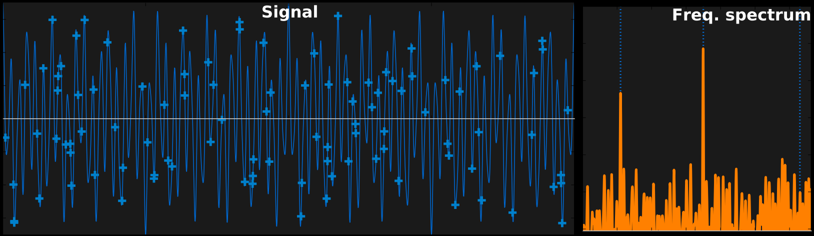


Solution: incoherent measurements



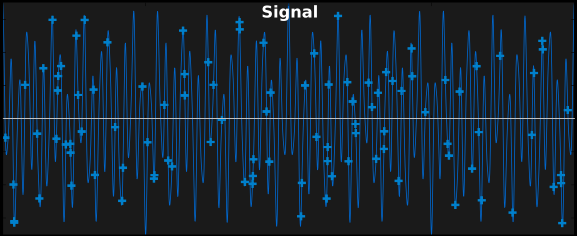
- i.e. careful choice of null-space of **A**

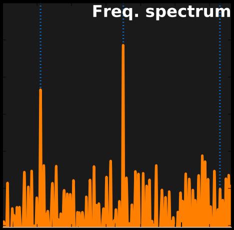
Incoherent measurements, but scarcity of data



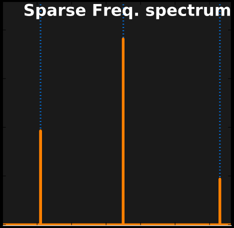
- The null-space of **A** is spread out in frequency
- Not much data $\Rightarrow$ large null-space
  $=$ captures "noise"

Incoherent measurements, but scarcity of data





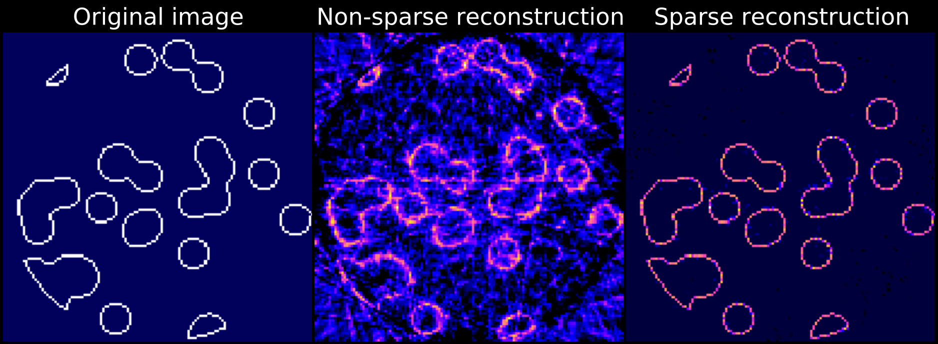
- The null-space of **A** is spread out in frequency
- Not much data $\Rightarrow$ large null-space
  $=$ captures "noise"



Impose sparsity

- Find a small number of frequencies to explain the signal

1 And for tomography reconstruction?

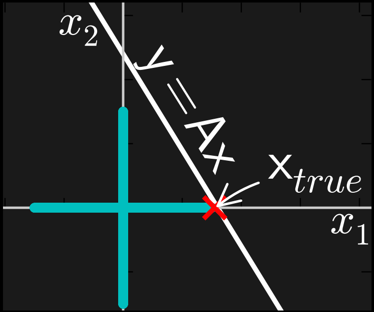


128×128 pixels, 18 projections

http://scikit-learn.org/stable/auto_examples/applications/plot_tomography_l1_reconstruction.html

1 Why does it work: a geometric explanation

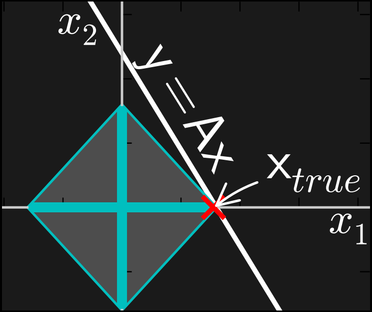
- Two coefficients of **x** not in the null-space of **A**:



- The sparsest solution is in the blue cross
- It corresponds to the true solution ($\mathbf{x}_{\text{true}}$) if the slope is $> 45°$

Why does it work: a geometric explanation

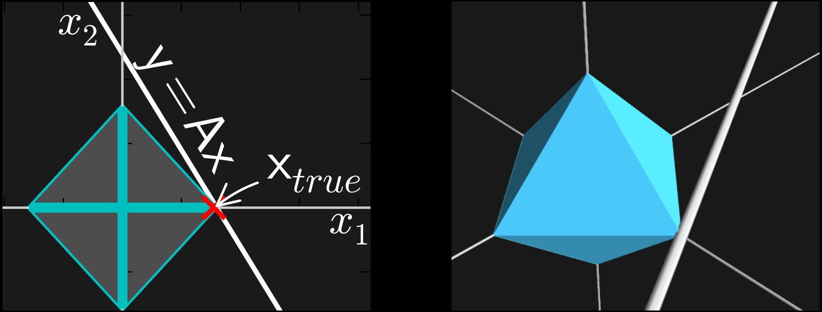
- Two coefficients of **x** not in the null-space of **A**:



- The sparsest solution is in the blue cross
- It corresponds to the true solution ($\mathbf{x}_{\text{true}}$) if the slope is $> 45°$
- The cross can be replaced by its convex hull

1 Why does it work: a geometric explanation

- Two coefficients of **x** not in the null-space of **A**:



- The sparsest solution is in the blue cross
- It corresponds to the true solution ($\mathbf{x}_{\text{true}}$) if the slope is $> 45°$
- In high dimension: large acceptable set

- Recovery of **sparse** signal

- Null space of sensing operator *incoherent* with sparse representation

$\Rightarrow$ **Excellent sparse recovery with little projections**

Minimum number of observations necessary:
$n_{\text{min}} \sim k \log p$, with k: number of non zeros

[Candes 2006]

Rmk Theory for *i.i.d.* samples
Related to *"compressive sensing"*

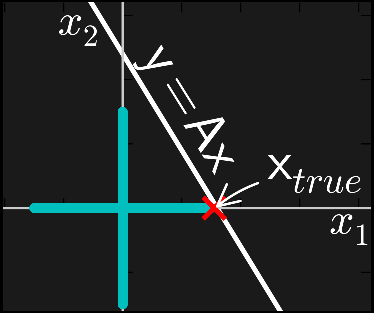
2 Mathematical formulation

- Variational formulation
- Introduction of noise

2 Maximizing the sparsity

- ℓ_0 number of non-zeros

$$\min_{\mathbf{x}} \ell_0(\mathbf{x}) \qquad s.t. \quad \mathbf{y} = \mathbf{A}\,\mathbf{x}$$



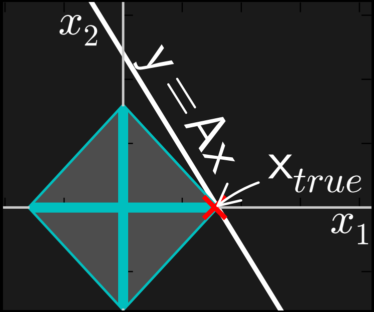
- "Matching pursuit" problem [Mallat, Zhang 1993]
  "Orthogonal matching pursuit" [Pati, *et al* 1993]

Problem: Non-convex optimization ☹

2 Maximizing the sparsity

- $\ell_1(\mathbf{x}) = \sum_i |\mathbf{x}_i|$

$$\min_{\mathbf{x}} \ell_1(\mathbf{x}) \qquad s.t. \quad \mathbf{y} = \mathbf{A}\,\mathbf{x}$$



- "Basis pursuit" [Chen, Donoho, Saunders 1998]

$$\mathbf{y} = \mathbf{A}\,\mathbf{x} + \mathbf{e} \qquad \mathbf{e} = \text{observation noise}$$

- New formulation:

$$\min_{\mathbf{x}} \ell_1(\mathbf{x}) \qquad s.t. \quad \cancel{\mathbf{y} = \mathbf{A}\,\mathbf{x}} \quad \|\mathbf{y} - \mathbf{A}\,\mathbf{x}\|_2^2 \leq \varepsilon^2$$

- Equivalent: “Lasso estimator” [Tibshirani 1996]

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\,\ell_1(\mathbf{x})$$

$$\mathbf{y} = \mathbf{A}\,\mathbf{x} + \mathbf{e} \qquad \mathbf{e} = \text{observation noise}$$

- New formulation:

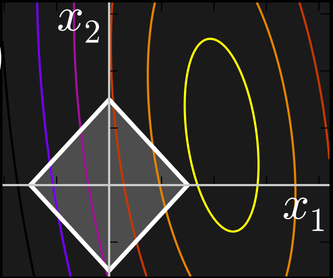
$$\min_{\mathbf{x}} \ell_1(\mathbf{x}) \qquad s.t. \quad \cancel{\mathbf{y} = \mathbf{A}\,\mathbf{x}} \quad \|\mathbf{y} - \mathbf{A}\,\mathbf{x}\|_2^2 \leq \varepsilon^2$$

- Equivalent: "Lasso estimator" [Tibshirani 1996]

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\,\ell_1(\mathbf{x})$$

Data fit

Penalization



Rmk: kink in the ℓ_1 ball creates sparsity

2 Probabilistic modeling: Bayesian interpretation

$$\mathcal{P}(\mathbf{x}|\mathbf{y}) \propto \mathcal{P}(\mathbf{y}|\mathbf{x})\ \mathcal{P}(\mathbf{x}) \qquad (\star)$$

"Posterior"
Quantity of interest

Forward model

"Prior"
Expectations on $\mathbf{x}$

- Forward model: $\mathbf{y} = \mathbf{A}\,\mathbf{x} + \mathbf{e}$, $\mathbf{e}$: Gaussian noise
  $\Rightarrow \mathcal{P}(\mathbf{y}|\mathbf{x}) \propto \exp -\frac{1}{2\,\sigma^2}\|\mathbf{y} - \mathbf{A}\,\mathbf{x}\|_2^2$

- Prior: Laplacian $\mathcal{P}(\mathbf{x}) \propto \exp -\frac{1}{\mu}\|\mathbf{x}\|_1$

Negated log of $(\star)$: $\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{A}\,\mathbf{x}\|_2^2 + \frac{1}{\mu}\ell_1(\mathbf{x})$

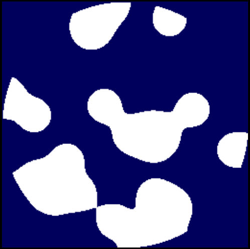
Maximum of posterior is Lasso estimate

Note that this picture is limited and the Lasso is not a good Bayesian estimator for the Laplace prior [Gribonval 2011].

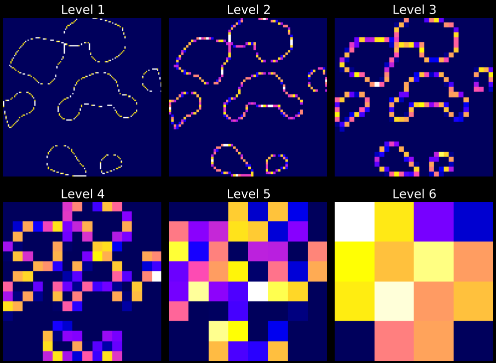
3 Choice of a sparse representation

- Sparse in wavelet domain
- Total variation

Typical images are not sparse



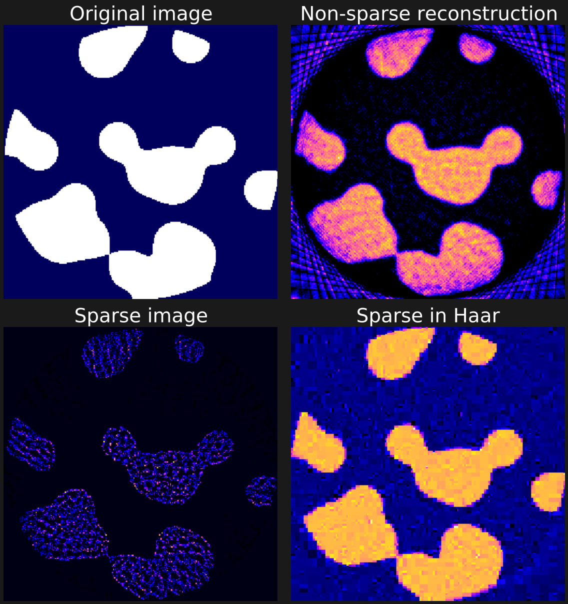
Haar decomposition



⇒ Impose sparsity in Haar representation

$\mathbf{A} \rightarrow \mathbf{A}\,\mathbf{H}$ where $\mathbf{H}$ is the Haar transform

3 Sparsity in wavelet representation

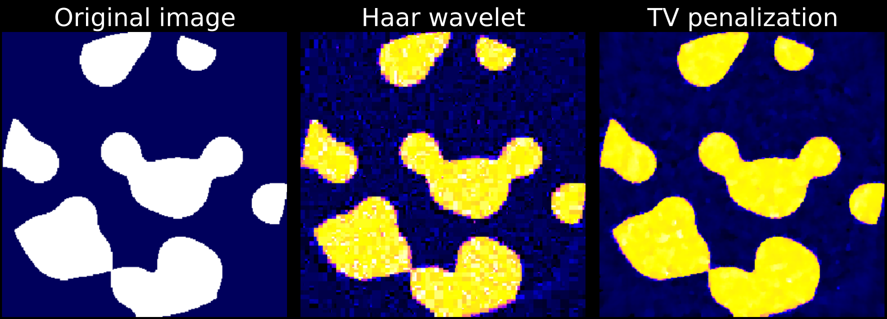


- Impose a sparse gradient

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \sum_i \|(\nabla \mathbf{x})_i\|_2$$

ℓ_{12} norm: ℓ_1 norm of the gradient magnitude

Sets ∇_x and ∇_y to zero jointly



3 Total variation

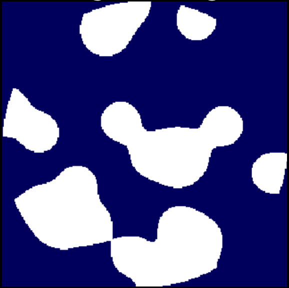
- Impose a sparse gradient

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \sum_i \big\|(\nabla \mathbf{x})_i\big\|_2$$

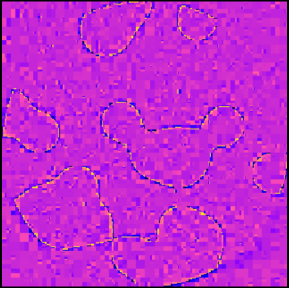
ℓ_{12} norm: ℓ_1 norm of the gradient magnitude

Sets ∇_x and ∇_y to zero jointly

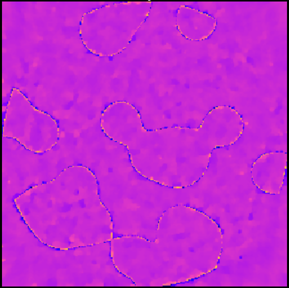
Original image



Error for Haar wavelet



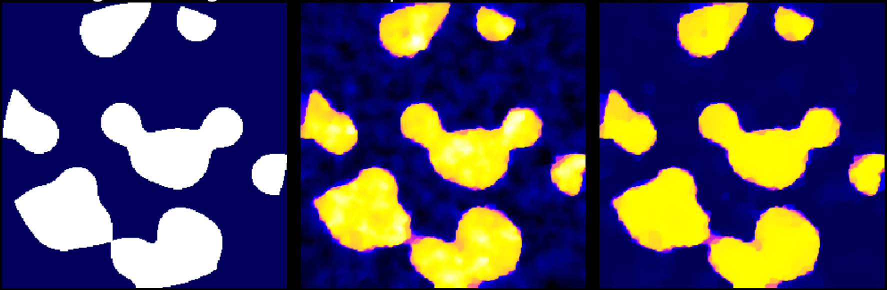
Error for TV penalization



- Bound $\mathbf{x}$ in $[0, 1]$

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \sum_i \left\|(\nabla \mathbf{x})_i\right\|_2 + \mathcal{I}([0, 1])$$

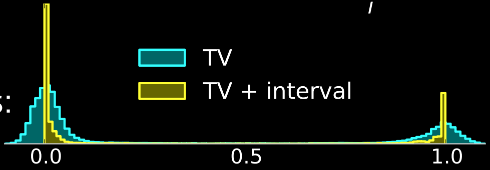
Original image | TV penalization | TV + interval



- Bound $\mathbf{x}$ in $[0, 1]$

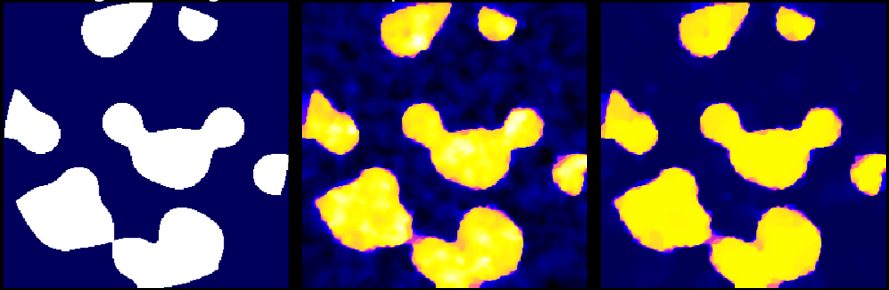
$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \sum_i \left\|(\nabla \mathbf{x})_i\right\|_2 + \mathcal{I}([0, 1])$$

Histograms:



Rmk: Constraint does more than folding values outside of the range back in.

Original image | TV penalization | TV + interval

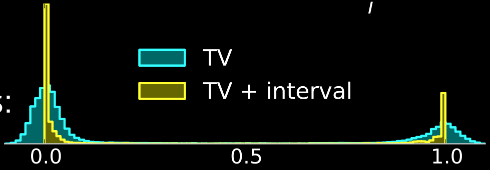


3 Total variation + interval

- Bound $\mathbf{x}$ in $[0, 1]$

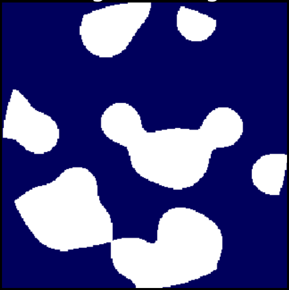
$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \sum_i \left\|(\nabla \mathbf{x})_i\right\|_2 + \mathcal{I}([0,1])$$

Histograms:

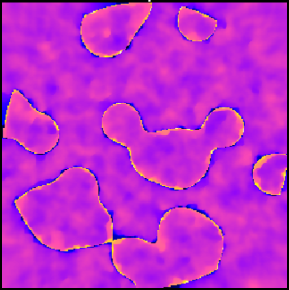


Rmk: Constraint does more than folding values outside of the range back in.

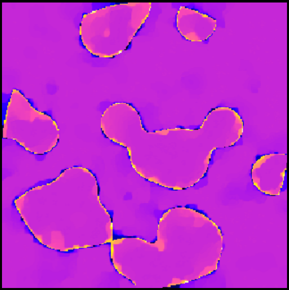
Original image



Error for TV penalization



Error for TV + interval



Analysis vs synthesis

- Wavelet basis $\min \|\mathbf{y} - \mathbf{A}\,\mathbf{H}\,\mathbf{x}\|_2^2 + \|\mathbf{x}\|_1$

  $\mathbf{H}$ Wavelet transform

- Total variation $\min \|\mathbf{y} - \mathbf{A}\,\mathbf{x}\|_2^2 + \|\mathbf{D}\mathbf{x}\|_1$

  $\mathbf{D}$ Spatial derivation operator (∇)

Analysis vs synthesis

- Wavelet basis $\quad \min \|\mathbf{y} - \mathbf{A}\,\mathbf{H}\,\mathbf{x}\|_2^2 + \|\mathbf{x}\|_1$

  H Wavelet transform

  "synthesis" formulation

- Total variation $\quad \min \|\mathbf{y} - \mathbf{A}\,\mathbf{x}\|_2^2 + \|\mathbf{D}\mathbf{x}\|_1$

  D Spatial derivation operator (∇)

  "analysis" formulation

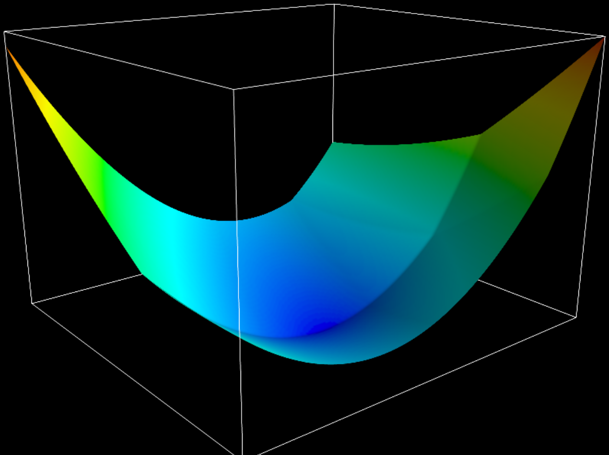
Theory and algorithms easier for synthesis

Equivalence *iif* **D** is invertible

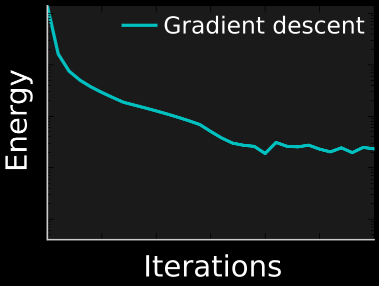
4 Optimization algorithms

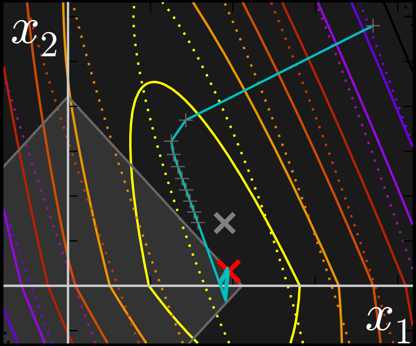
- Non-smooth optimization

  $\Rightarrow$ "proximal operators"



Gradient descent





- Smooth optimization fails in non-smooth regions
- These are specifically the spots that interest us

4 Iterative Shrinkage-Thresholding Algorithm

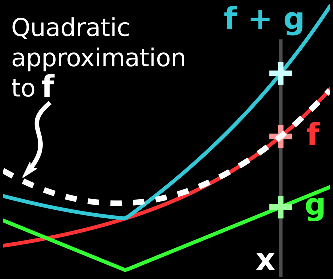
- Settings: $\min f + g$; f smooth, g non-smooth
  f and g convex, ∇f L-Lipschitz

- Typically f is the data fit term, and g the penalty

  ex: Lasso $\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{A}\,\mathbf{x}\|_2^2 + \frac{1}{\mu}\ell_1(\mathbf{x})$

- Settings: $\min f + g$; f smooth, g non-smooth
  f and g convex, ∇f L-Lipschitz

- Minimize successively:
  (quadratic approx of f) $+ g$



$$f(\mathbf{x}) < f(\mathbf{y}) + \left\langle \mathbf{x} - \mathbf{y}, \nabla f(\mathbf{y}) \right\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$$

Proof: ▪ by convexity $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{y})\,(\mathbf{y} - \mathbf{x})$
▪ in the second term: $\nabla f(\mathbf{y}) \rightarrow \nabla f(\mathbf{x}) + (\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))$
▪ upper bound last term with Lipschitz continuity of ∇f

$$\mathbf{x}_{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \left(g(\mathbf{x}) + \frac{L}{2}\left\| \mathbf{x} - \left(\mathbf{x_k} - \frac{1}{L}\nabla f(\mathbf{x_k})\right)\right\|_2^2 \right)$$

[Daubechies 2004]

4 Iterative Shrinkage-Thresholding Algorithm

Step 1: Gradient descent on f

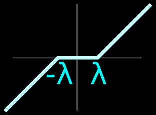
Step 2: Proximal operator of g:

$$\text{prox}_{\lambda g}(\mathbf{x}) \overset{def}{=} \underset{\mathbf{y}}{\text{argmin}}\ \|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda\, g(\mathbf{y})$$

- Generalization of Euclidean projection on convex set $\{\mathbf{x},\ g(\mathbf{x}) \leq 1\}$

Rmk: if g is the indicator function of a set $\mathcal{S}$, the proximal operator is the Euclidean projection.

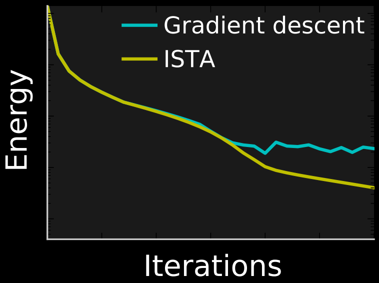
- $\text{prox}_{\lambda \ell_1}(\mathbf{x}) = \text{sign}(\mathbf{x}_i)\big(\mathbf{x}_i - \lambda\big)_+$
  "soft thresholding"

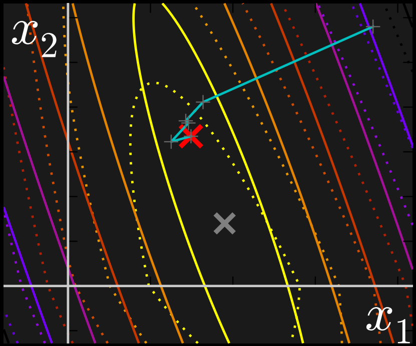


$$\mathbf{x}_{k+1} = \underset{\mathbf{x}}{\text{argmin}} \left(g(\mathbf{x}) + \frac{L}{2}\left\|\mathbf{x} - \left(\mathbf{x_k} - \frac{1}{L}\nabla f(\mathbf{x_k})\right)\right\|_2^2 \right)$$

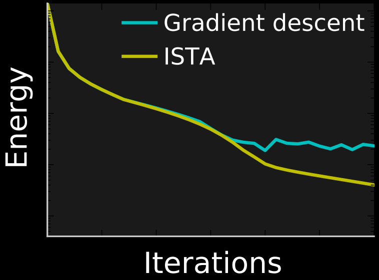
[Daubechies 2004]

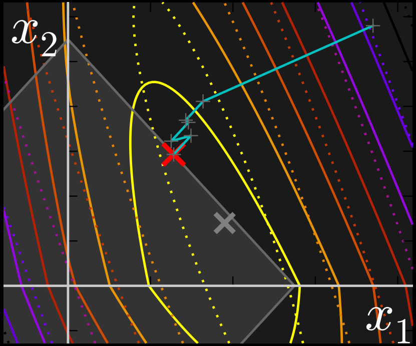
4 Iterative Shrinkage-Thresholding Algorithm





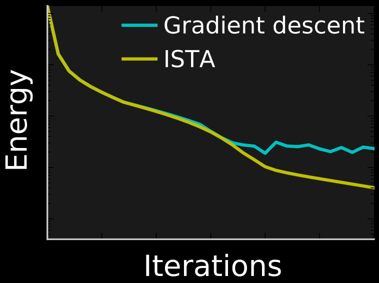
Gradient descent step

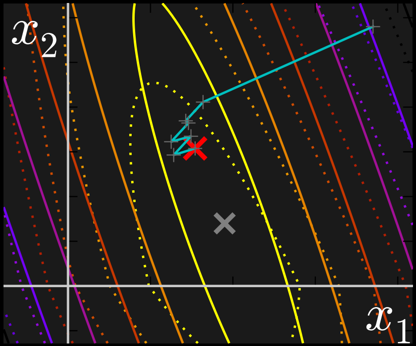




Projection on ℓ_1 ball

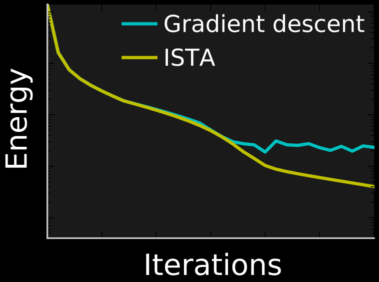
4 Iterative Shrinkage-Thresholding Algorithm

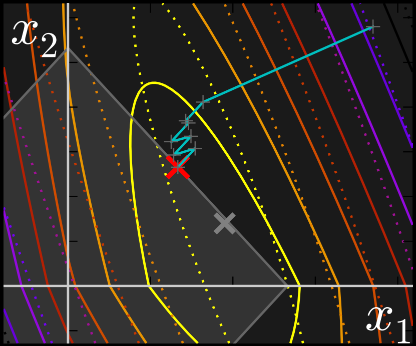




Gradient descent step

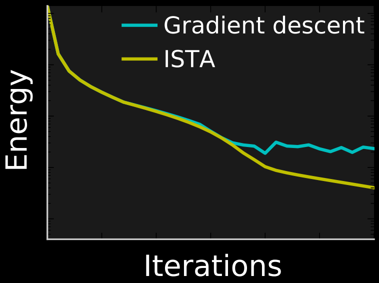
4 Iterative Shrinkage-Thresholding Algorithm

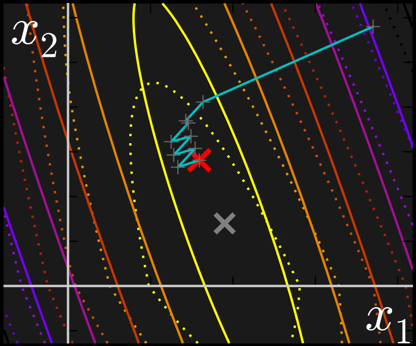




Projection on ℓ_1 ball

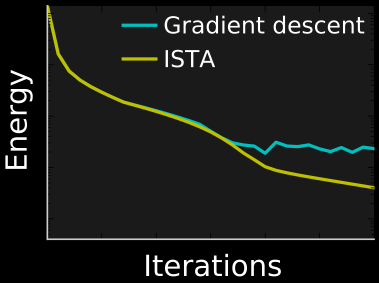
4 Iterative Shrinkage-Thresholding Algorithm

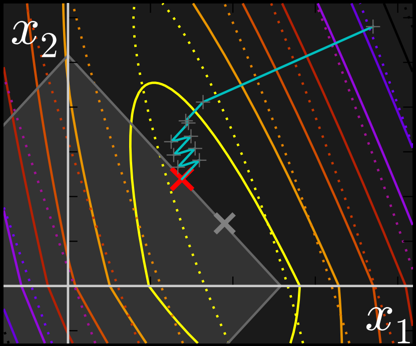




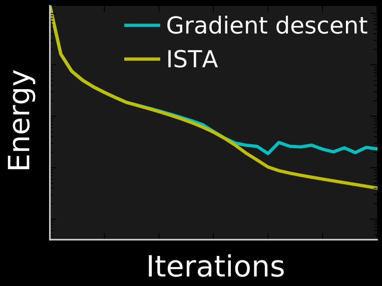
Gradient descent step

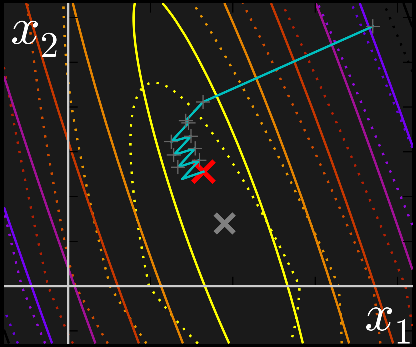
4 Iterative Shrinkage-Thresholding Algorithm



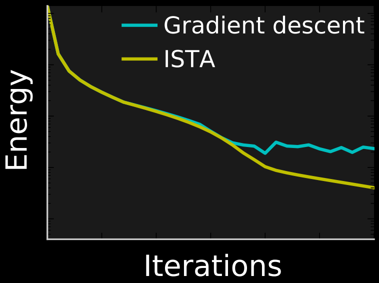


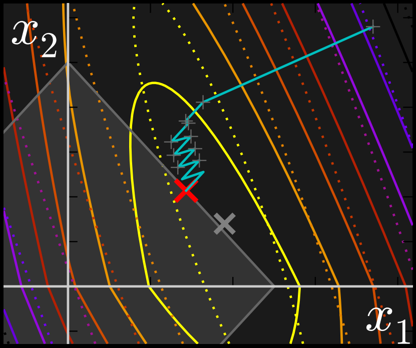
Projection on ℓ_1 ball





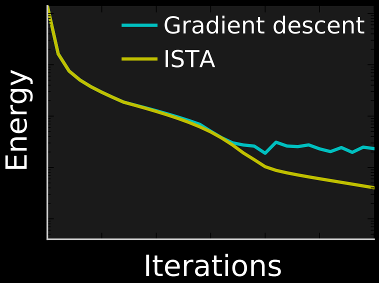
Gradient descent step

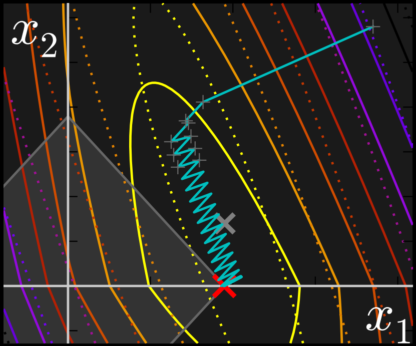


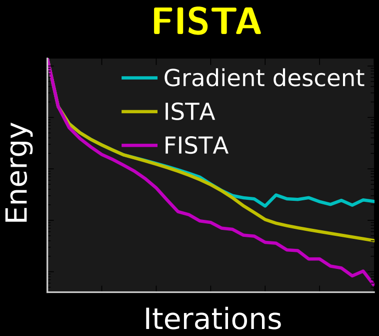


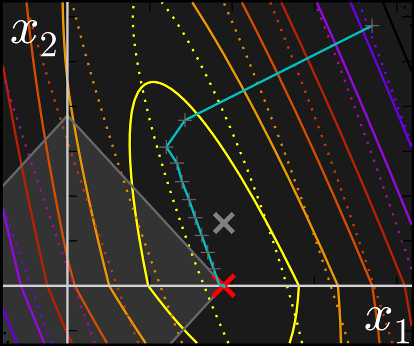
Projection on ℓ_1 ball

4 Iterative Shrinkage-Thresholding Algorithm









- As with conjugate gradient: add a memory term

- $d\mathbf{x}_{k+1} = d\mathbf{x}_{k+1}^{ISTA} + \frac{t_k - 1}{t_{k+1}}(d\mathbf{x}_k - d\mathbf{x}_{k-1})$

$$t_1 = 1, \; t_{k+1} = \frac{1+\sqrt{1+4\,t_k^2}}{2}$$

$\Rightarrow \mathcal{O}(k^{-2})$ convergence

[Beck Teboulle 2009]

4 Proximal operator for total variation

Reformulate to smooth + non-smooth with a simple projection step and use FISTA: [Chambolle 2004]

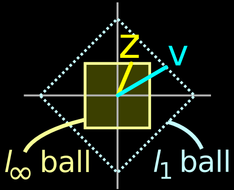
$$\mathrm{prox}_{\lambda \mathrm{TV}}\mathbf{x} \;=\; \underset{\mathbf{x}}{\mathrm{argmin}}\, \|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda \sum_i \left\|(\nabla \mathbf{x})_i\right\|_2$$

4 Proximal operator for total variation

Reformulate to smooth + non-smooth with a simple projection step and use FISTA: [Chambolle 2004]

$$\begin{aligned}\mathrm{prox}_{\lambda \mathrm{TV}} \mathbf{x} &= \underset{\mathbf{x}}{\mathrm{argmin}} \, \|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda \sum_i \left\|(\nabla \mathbf{x})_i\right\|_2 \\ &= \underset{\mathbf{z},\, \|\mathbf{z}\|_\infty \leq 1}{\mathrm{argmax}} \, \|\lambda \,\mathrm{div}\, \mathbf{z} + \mathbf{y}\|_2^2\end{aligned}$$

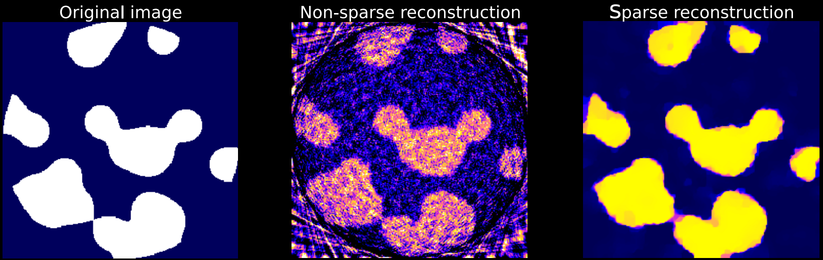
Proof:



- "dual norm": $\|\mathbf{v}\|_1 = \max_{\|\mathbf{z}\|_\infty \leq 1} \langle \mathbf{v}, \mathbf{z} \rangle$
- div is the adjoint of ∇: $\langle \nabla \mathbf{v}, \mathbf{z} \rangle = \langle \mathbf{v}, -\mathrm{div}\, \mathbf{z} \rangle$
- Swap min and max and solve for $\mathbf{x}$

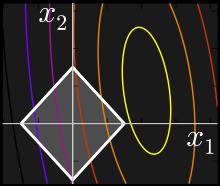
Duality: [Boyd 2004] This proof: [Michel 2011]

Sparsity for compressed tomography reconstruction



Sparsity for compressed tomography reconstruction

- Add penalizations with kinks
- Choice of prior/sparse representation
- Non-smooth optimization (FISTA)



Further discussion: choice of prior/parameters

- Minimize reconstruction error from degraded data of gold-standard acquisitions
- Cross-validation: leave half of the projections and minimize projection error of reconstruction

Python code available:
`https://github.com/emmanuelle/tomo-tv`

- [Candes 2006] E. Candès, J. Romberg and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, Trans Inf Theory, (52) 2006

- [Wainwright 2009] M. Wainwright, *Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Usingℓ_1 constrained quadratic programming (Lasso)*, Trans Inf Theory, (55) 2009

- [Mallat, Zhang 1993] S. Mallat and Z. Zhang, *Matching pursuits with Time-Frequency dictionaries*, Trans Sign Proc (41) 1993

- [Pati, *et al* 1993] Y. Pati, R. Rezaiifar, P. Krishnaprasad, *Orthogonal matching pursuit: Recursive function approximation with plications to wavelet decomposition*, 27th Signals, Systems and Computers Conf 1993

Bibliography (2/3)

- [Chen, Donoho, Saunders 1998] S. Chen, D. Donoho, M. Saunders, *Atomic decomposition by basis pursuit*, SIAM J Sci Computing (20) 1998

- [Tibshirani 1996] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J Roy Stat Soc B, 1996

- [Gribonval 2011] R. Gribonval, *Should penalized least squares regression be interpreted as Maximum A Posteriori estimation?*, Trans Sig Proc, (59) 2011

- [Daubechies 2004] I. Daubechies, M. Defrise, C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm Pure Appl Math, (57) 2004

Bibliography (2/3)

- [Beck Teboulle 2009], A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J Imaging Sciences, (2) 2009

- [Chambolle 2004], A. Chambolle, *An algorithm for total variation minimization and applications*, J Math imag vision, (20) 2004

- [Boyd 2004], S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press 2004
  — **Reference on convex optimization and duality**

- [Michel 2011], V. Michel *et al.*, *Total variation regularization for fMRI-based prediction of behaviour*, Trans Med Imag (30) 2011
  — **Proof of TV reformulation: appendix C**