Sparse Estimation for Image and Vision Processing

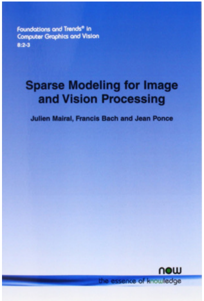
Julien Mairal

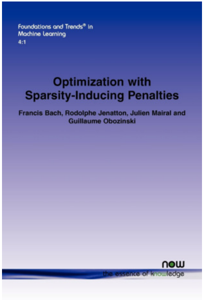
Inria, Grenoble

Ohrid, September 2016

Course material (freely available on arXiv)

J. Mairal, F. Bach and J. Ponce. *Sparse Modeling for Image and Vision Processing*. Foundations and Trends in Computer Graphics and Vision. 2014.



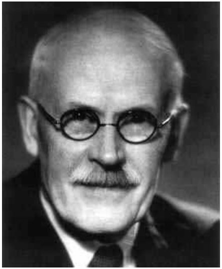


F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. *Optimization with sparsity-inducing penalties*. Foundations and Trends in Machine Learning, 4(1). 2012.

Part I: A Short Introduction to Parcimony

Early thoughts

(a) Dorothy Wrinch
1894–1980



(b) Harold Jeffreys
1891–1989

The existence of simple laws is, then, apparently, to be regarded as a quality of nature; and accordingly we may infer that it is justifiable to prefer a simple law to a more complex one that fits our observations slightly better.

[Wrinch and Jeffreys, 1921]. Philosophical Magazine Series.

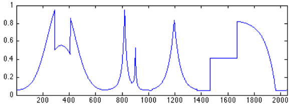
Historical overview of parsimony

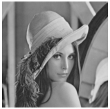
- 14th century: Ockham's razor;
- 1921: Wrinch and Jeffreys' simplicity principle;
- 1952: Markowitz's portfolio selection;
- 60 and 70's: best subset selection in statistics;
- 70's: **use of the ℓ_1-norm** for signal recovery in geophysics;
- 90's: wavelet thresholding in signal processing;
- 1996: Olshausen and Field's **dictionary learning**;
- 1996–1999: Lasso (statistics) and basis pursuit (signal processing);
- 2006: compressed sensing (signal processing) and Lasso consistency (statistics);
- 2006–now: **applications of dictionary learning in various scientific fields such as image processing and computer vision**.

The modern parsimony and the ℓ_1-norm

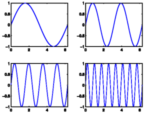
Sparse linear models in signal processing

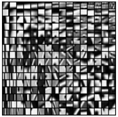
Let $\mathbf{x}$ in $\mathbb{R}^n$ be a signal.





Let $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_p] \in \mathbb{R}^{n \times p}$ be a set of elementary signals.
We call it **dictionary**.





$\mathbf{D}$ is "adapted" to $\mathbf{x}$ if it can represent it with a few elements—that is, there exists a **sparse vector** $\boldsymbol{\alpha}$ in $\mathbb{R}^p$ such that $\mathbf{x} \approx \mathbf{D}\boldsymbol{\alpha}$. We call $\boldsymbol{\alpha}$ the **sparse code**.

$$\underbrace{\begin{pmatrix} \\ \mathbf{x} \\ \\ \end{pmatrix}}_{\mathbf{x}\in\mathbb{R}^n} \approx \underbrace{\left(\begin{array}{c|c|c|c} & & & \\ \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_p \\ & & & \end{array}\right)}_{\mathbf{D}\in\mathbb{R}^{n\times p}} \underbrace{\begin{pmatrix} \boldsymbol{\alpha}[1] \\ \boldsymbol{\alpha}[2] \\ \vdots \\ \boldsymbol{\alpha}[p] \end{pmatrix}}_{\boldsymbol{\alpha}\in\mathbb{R}^p,\textbf{sparse}}$$

How do we find $\boldsymbol{\alpha}$?

We try to solve the sparse approximation problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \text{ s.t. } \|\boldsymbol{\alpha}\|_0 \leq k,$$

but...

The modern parsimony and the ℓ_1-norm

Sparse linear models in signal processing

How do we find $\boldsymbol{\alpha}$?

We try to solve the sparse approximation problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \;\text{ s.t. }\; \|\boldsymbol{\alpha}\|_0 \leq k,$$

but... **the problem is NP-hard [Natarajan, 1995].**

Strategy 1: try anyway

use greedy algorithm to find an approximate solution.

Strategy 2: use a convex relaxation

replace ℓ_0 by ℓ_1.

The modern parsimony and the ℓ_1-norm

Sparse linear models: machine learning/statistics point of view

Let $(y_i, \mathbf{x}_i)_{i=1}^n$ be a training set, where the vectors $\mathbf{x}_i$ are in $\mathbb{R}^p$ and are called features. The scalars y_i are in

- $\{-1, +1\}$ for **binary** classification problems.
- $\mathbb{R}$ for **regression** problems.

We assume there exists a relation $y \approx \boldsymbol{\beta}^\top \mathbf{x}$, and solve

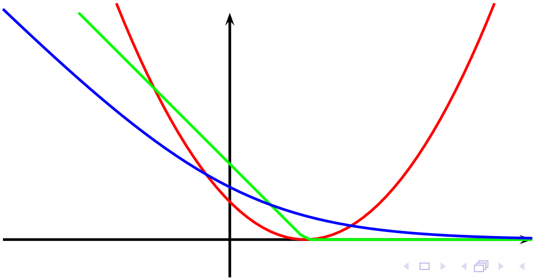
$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \underbrace{\frac{1}{n} \sum_{i=1}^{n} L(y_i, \boldsymbol{\beta}^\top \mathbf{x}_i)}_{\text{empirical risk}} + \underbrace{\lambda \psi(\boldsymbol{\beta})}_{\text{regularization}} .$$

A few examples:

Ridge regression:
$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 + \lambda\|\boldsymbol{\beta}\|_2^2.$$

Linear SVM:
$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i\boldsymbol{\beta}^\top \mathbf{x}_i) + \lambda\|\boldsymbol{\beta}\|_2^2.$$

Logistic regression:
$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i\boldsymbol{\beta}^\top \mathbf{x}_i}\right) + \lambda\|\boldsymbol{\beta}\|_2^2.$$



The modern parsimony and the ℓ_1-norm

Sparse linear models: machine learning/statistics point of view

A few examples:

Ridge regression:
$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 + \lambda\|\boldsymbol{\beta}\|_2^2.$$

Linear SVM:
$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i\boldsymbol{\beta}^\top \mathbf{x}_i) + \lambda\|\boldsymbol{\beta}\|_2^2.$$

Logistic regression:
$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i\boldsymbol{\beta}^\top \mathbf{x}_i}\right) + \lambda\|\boldsymbol{\beta}\|_2^2.$$

The **squared ℓ_2-norm** induces "**smoothness**" in $\boldsymbol{\beta}$. When one knows in advance that $\boldsymbol{\beta}$ should be sparse, one should use a **sparsity-inducing** regularization such as the **ℓ_1-norm**. [Chen et al., 1999, Tibshirani, 1996]

The modern parsimony and the ℓ_1-norm

Originally used to induce sparsity in geophysics [Claerbout and Muir, 1973, Taylor et al., 1979], the ℓ_1-norm became popular in statistics with the **Lasso** [Tibshirani, 1996] and in signal processing with the **Basis pursuit** [Chen et al., 1999].

Three "equivalent" formulations

1. $$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1;$$

2. $$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\alpha}\|_1 \leq \mu;$$

3. $$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\boldsymbol{\alpha}\|_1 \ \text{ s.t. } \ \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq \varepsilon.$$

The modern parsimony and the ℓ_1-norm

And some variants...

For noiseless problems

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\boldsymbol{\alpha}\|_1 \ \text{ s.t. } \ \mathbf{x} = \mathbf{D}\boldsymbol{\alpha}.$$

Beyond least squares

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} f(\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|_1,$$

where $f : \mathbb{R}^p \to \mathbb{R}$ is convex.

The modern parsimony and the ℓ_1-norm

And some variants...

For noiseless problems

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\boldsymbol{\alpha}\|_1 \;\; \text{s.t.} \;\; \mathbf{x} = \mathbf{D}\boldsymbol{\alpha}.$$

Beyond least squares

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} f(\boldsymbol{\alpha}) + \lambda\|\boldsymbol{\alpha}\|_1,$$

where $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is convex.

An important question remains:

why does the ℓ_1-norm induce sparsity?

The modern parsimony and the ℓ_1-norm

Why does the ℓ_1-norm induce sparsity?

Can we get some intuition from the simplest isotropic case?

$$\hat{\boldsymbol{\alpha}}(\lambda) = \operatorname*{arg\,min}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1,$$

or equivalently the Euclidean projection onto the ℓ_1-ball?

$$\tilde{\boldsymbol{\alpha}}(\mu) = \operatorname*{arg\,min}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \boldsymbol{\alpha}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\alpha}\|_1 \leq \mu.$$

"equivalent" means that for all $\lambda > 0$, there exists $\mu \geq 0$ such that $\tilde{\boldsymbol{\alpha}}(\mu) = \hat{\boldsymbol{\alpha}}(\lambda)$.

The modern parsimony and the ℓ_1-norm

Why does the ℓ_1-norm induce sparsity?

Can we get some intuition from the simplest isotropic case?

$$\hat{\boldsymbol{\alpha}}(\lambda) = \underset{\boldsymbol{\alpha} \in \mathbb{R}^p}{\arg\min} \frac{1}{2}\|\mathbf{x} - \boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1,$$

or equivalently the Euclidean projection onto the ℓ_1-ball?

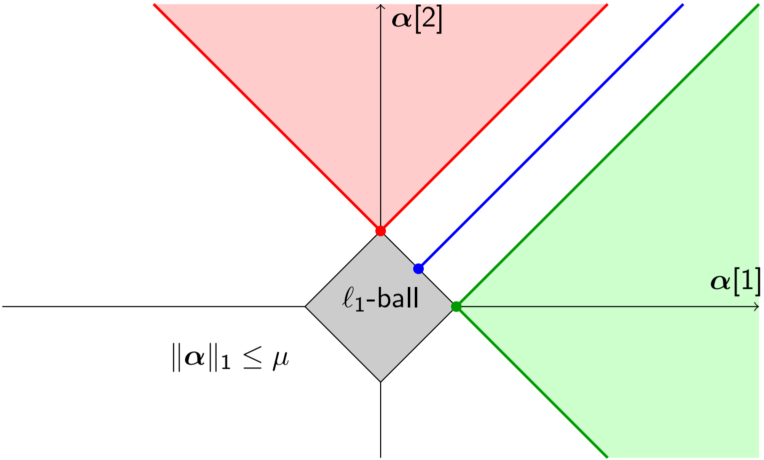
$$\tilde{\boldsymbol{\alpha}}(\mu) = \underset{\boldsymbol{\alpha} \in \mathbb{R}^p}{\arg\min} \frac{1}{2}\|\mathbf{x} - \boldsymbol{\alpha}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\alpha}\|_1 \leq \mu.$$

"equivalent" means that for all $\lambda > 0$, there exists $\mu \geq 0$ such that $\tilde{\boldsymbol{\alpha}}(\mu) = \hat{\boldsymbol{\alpha}}(\lambda)$.

The relation between μ and λ is unknown a priori.

Why does the ℓ_1-norm induce sparsity?

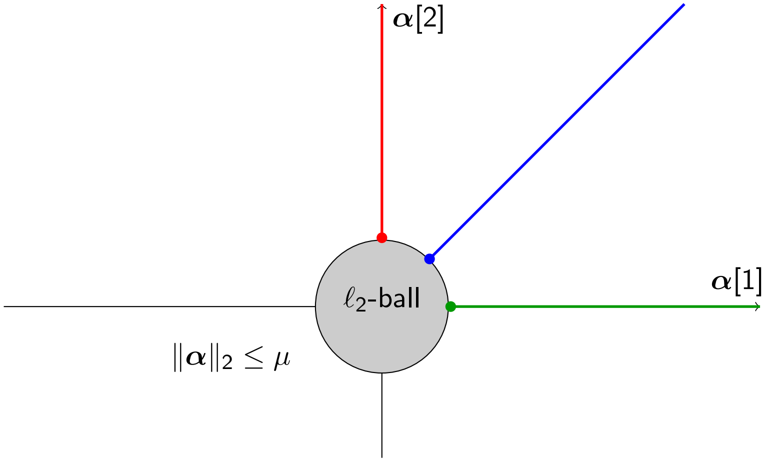
Regularizing with the ℓ_1-norm



The projection onto a convex set is "biased" towards singularities.

Why does the ℓ_1-norm induce sparsity?

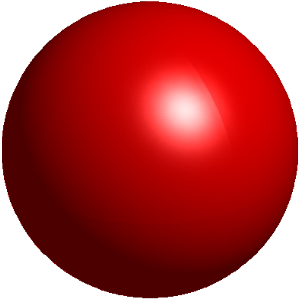
Regularizing with the ℓ_2-norm

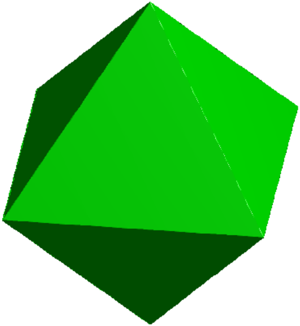


The ℓ_2-norm is isotropic.

Why does the ℓ_1-norm induce sparsity?

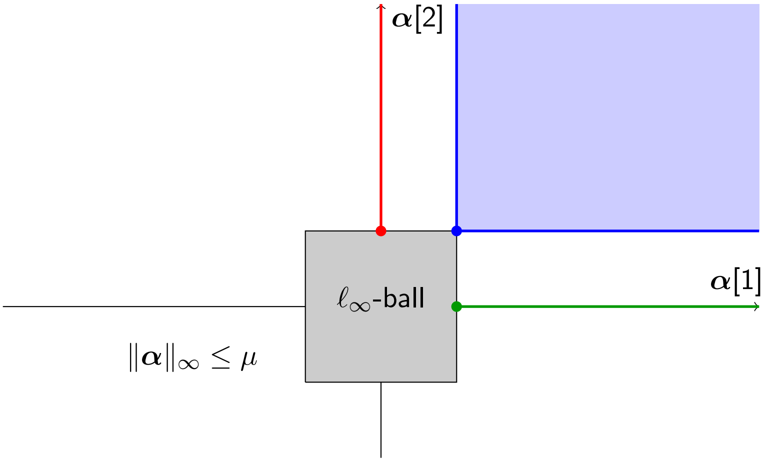
In 3D. (images produced by G. Obozinski)





Why does the ℓ_1-norm induce sparsity?

Regularizing with the ℓ_∞-norm



The ℓ_∞-norm encourages $|\boldsymbol{\alpha}[1]| = |\boldsymbol{\alpha}[2]|$.

Why does the ℓ_1-norm induce sparsity?

Analytical point of view: 1D case

$$\min_{\alpha \in \mathbb{R}} \frac{1}{2}(x-\alpha)^2 + \lambda|\alpha|$$

Piecewise quadratic function with a kink at zero.

Derivative at 0_+: $g_+ = -x + \lambda$ and 0_-: $g_- = -x - \lambda$.

Optimality conditions. α is optimal iff:

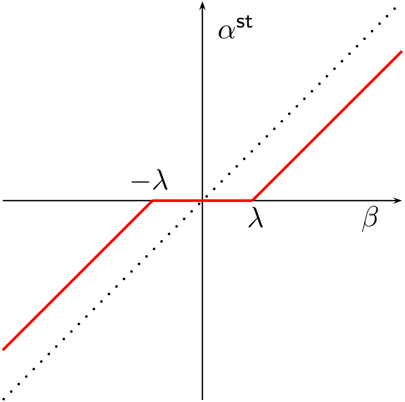
- $|\alpha| > 0$ and $(x - \alpha) + \lambda \,\text{sign}(\alpha) = 0$
- $\alpha = 0$ and $g_+ \geq 0$ and $g_- \leq 0$

The solution is a **soft-thresholding**:

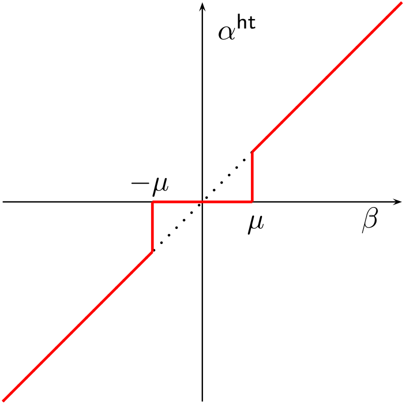
$$\alpha^\star = \text{sign}(x)(|x| - \lambda)^+.$$

Why does the ℓ_1-norm induce sparsity?

Analytical point of view: 1D case



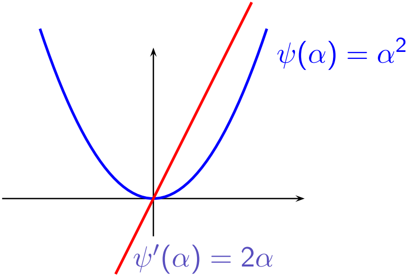
(c) Soft-thresholding operator,
$\alpha^{\text{st}} = \text{sign}(\beta)\max(|\beta| - \lambda, 0)$.

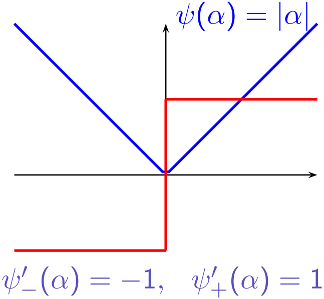


(d) Hard-thresholding operator
$\alpha^{\text{ht}} = \delta_{|\beta| \geq \mu}\beta$.

Why does the ℓ_1-norm induce sparsity?

Comparison with ℓ_2-regularization in 1D

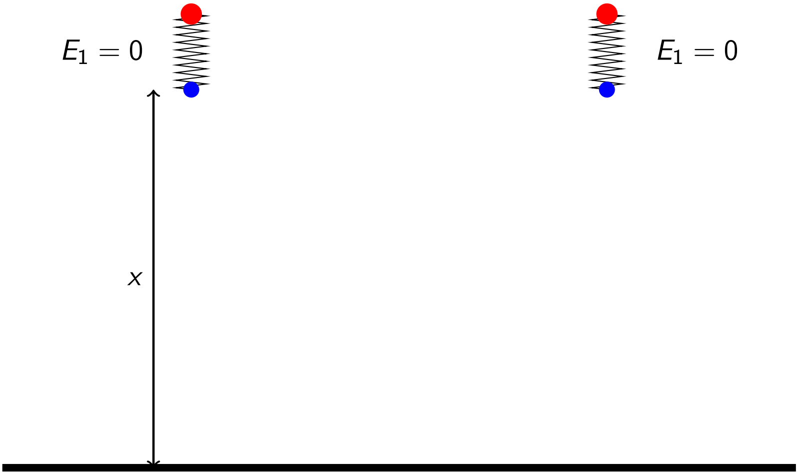




The gradient of the ℓ_2-penalty vanishes when α get close to 0. On its differentiable part, the norm of the gradient of the ℓ_1-norm is constant.

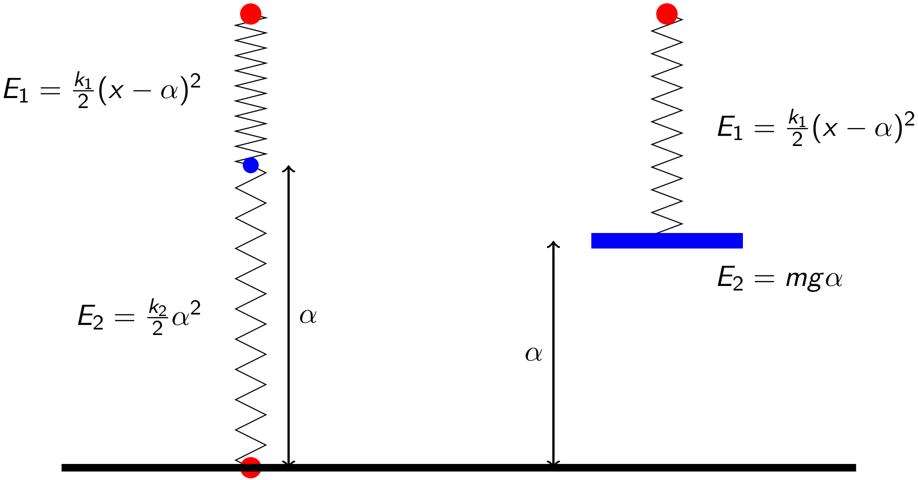
Why does the ℓ_1-norm induce sparsity?

Physical illustration



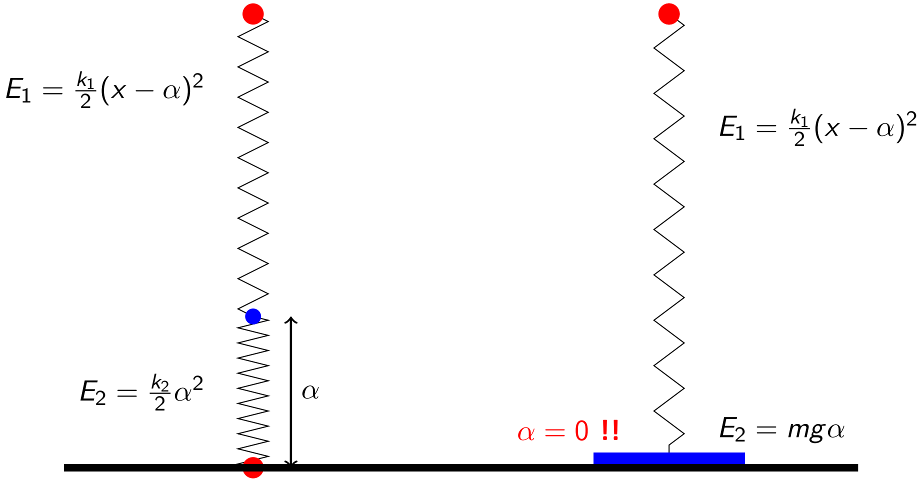
Why does the ℓ_1-norm induce sparsity?

Physical illustration

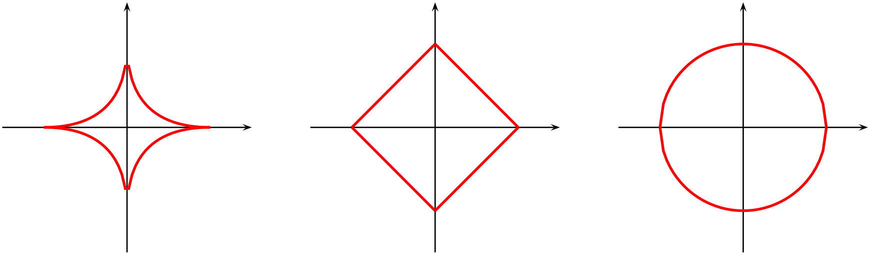


Why does the ℓ_1-norm induce sparsity?

Physical illustration



Non-convex sparsity-inducing penalties

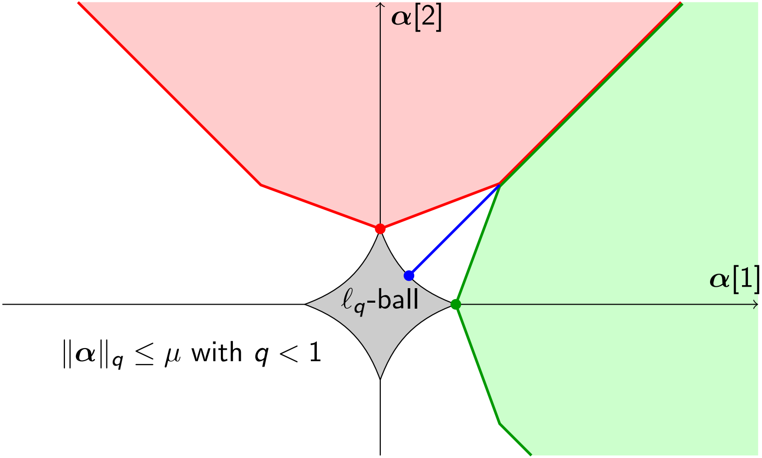


(e) $\ell_{0.5}$-ball, 2-D (f) ℓ_1-ball, 2-D (g) ℓ_2-ball, 2-D

Figure: Open balls in 2-D corresponding to several ℓ_q-norms and pseudo-norms.

$$\|\boldsymbol{\alpha}\|_q^q = \sum_{j=1}^{p} |\boldsymbol{\alpha}[j]|^q.$$

Non-convex sparsity-inducing penalties

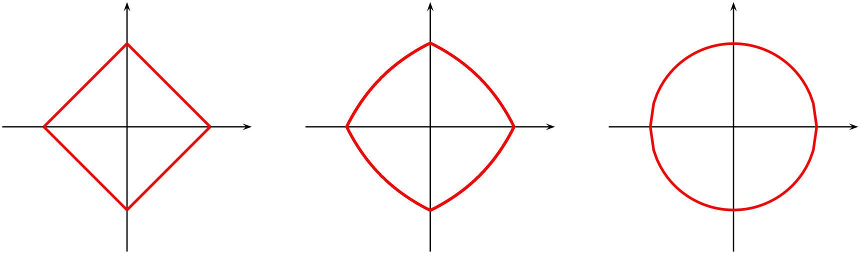


Elastic-net

The **elastic net** introduced by [Zou and Hastie, 2005]

$$\psi(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1 + \gamma\|\boldsymbol{\alpha}\|_2^2,$$

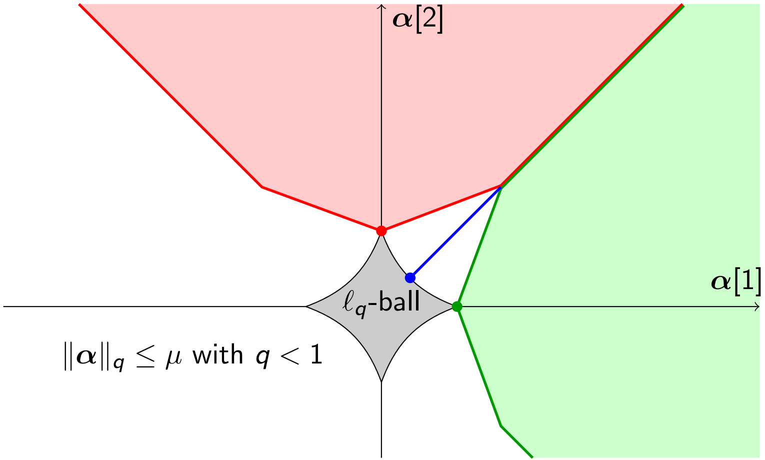
The penalty provides more stable (but less sparse) solutions.



(a) ℓ_1-ball, 2-D (b) elastic-net, 2-D (c) ℓ_2-ball, 2-D

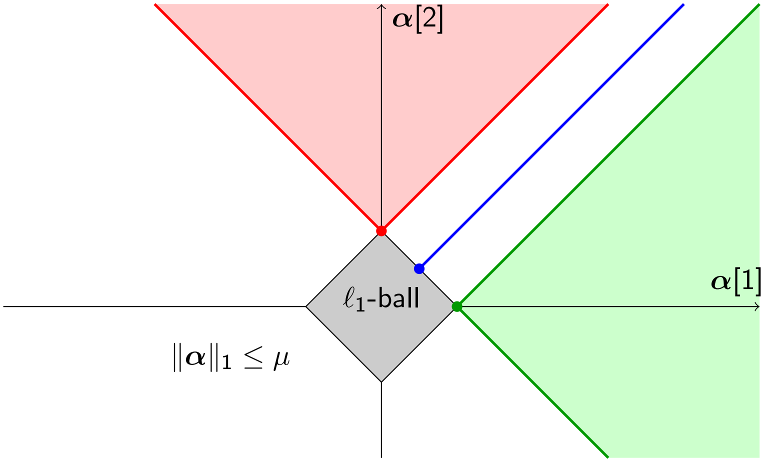
The elastic-net

vs other penalties



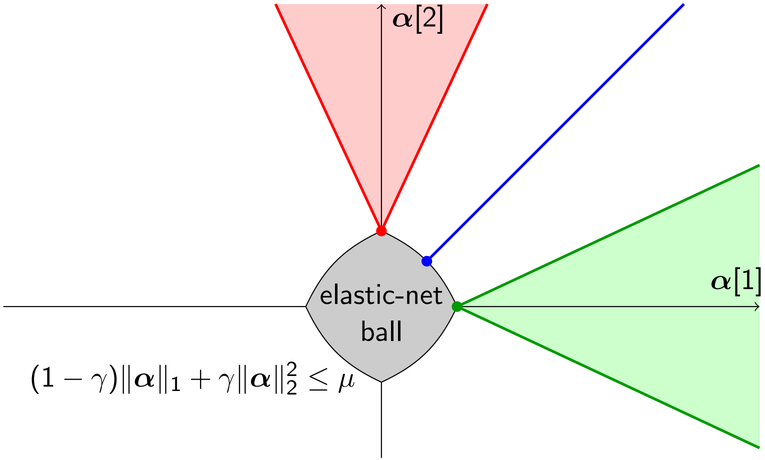
The elastic-net

vs other penalties



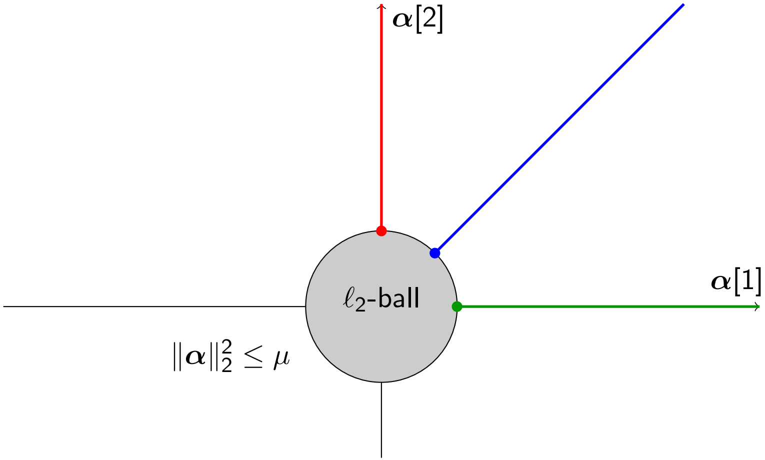
The elastic-net

vs other penalties

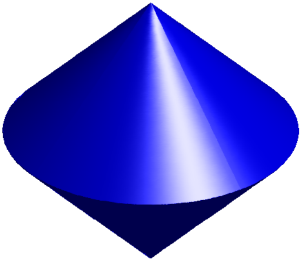


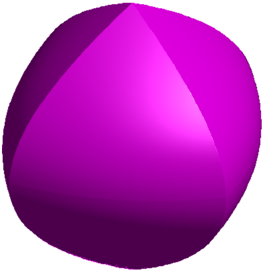
The elastic-net

vs other penalties

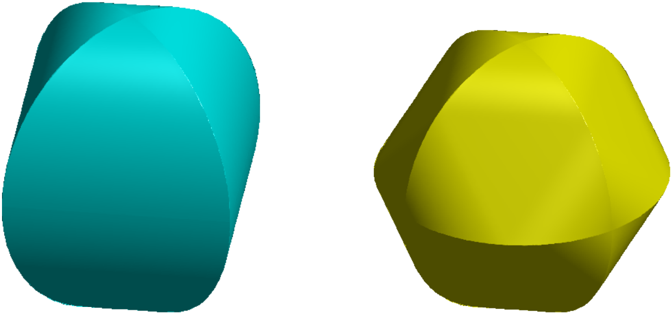


Structured sparsity





Structured sparsity



Part II: Discovering the structure of natural images

Dictionary learning

Neuroscientists were the first to automatically learn local structures in natural images.

The model of Olshausen and Field [1996] looks for a dictionary $\mathbf{D}$ adapted to a training set of natural image patches $\mathbf{x}_i$, $i = 1, \ldots, n$:

$$\min_{\mathbf{D} \in \mathcal{C}, \mathbf{A} \in \mathbb{R}^{p \times n}} \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \psi(\boldsymbol{\alpha}_i),$$

where $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]$ and $\mathcal{C} \triangleq \{\mathbf{D} \in \mathbb{R}^{m \times p} : \forall\, j,\ \|\mathbf{d}_j\|_2 \leq 1\}$.

Typical settings

- $n \approx 100\,000$;
- $m = 10 \times 10$ pixels;
- $p = 256$.

Dictionary learning

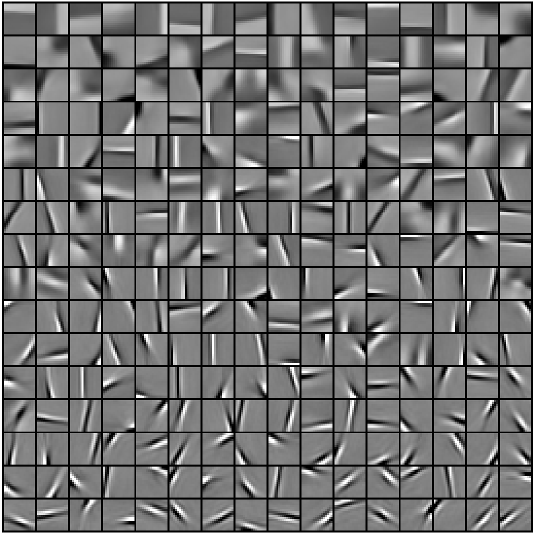


Figure: with centering

Dictionary learning

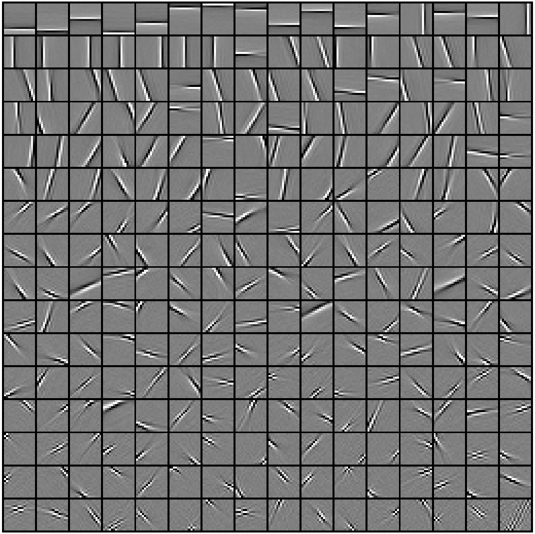


Figure: with whitening

Dictionary learning

Why was it found impressive by neuroscientists?

- since Hubel and Wiesel [1968], it is known that some visual neurons are responding to particular image features, such as oriented edges.
- Later, Daugman [1985] demonstrated that fitting a linear model to neuronal responses given a visual stimuli may produce filters that can be well approximated by a two-dimensional Gabor function.
- the original motivation of Olshausen and Field [1996] was to establish a relation between the statistical structure of natural images and the properties of neurons from area V1.

The results provided some "support" for classical models of V1 based on Gabor filters.

Dictionary learning

Why was it found impressive by neuroscientists?

- since Hubel and Wiesel [1968], it is known that some visual neurons are responding to particular image features, such as oriented edges.
- Later, Daugman [1985] demonstrated that fitting a linear model to neuronal responses given a visual stimuli may produce filters that can be well approximated by a two-dimensional Gabor function.
- the original motivation of Olshausen and Field [1996] was to establish a relation between the statistical structure of natural images and the properties of neurons from area V1.

Warning

In fact, little is known about the early visual cortex [Olshausen and Field, 2005, Carandini et al., 2005].

Matrix factorization

It is useful to see dictionary learning as a matrix factorization problem

$$\min_{\mathbf{D}\in\mathcal{C},\mathbf{A}\in\mathbb{R}^{p\times n}} \frac{1}{2n}\|\mathbf{X}-\mathbf{DA}\|_F^2 + \lambda\Psi(\mathbf{A}).$$

This is simply a matter of notation:

$$\min_{\mathbf{D}\in\mathcal{C},\mathbf{A}\in\mathbb{R}^{p\times n}} \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i-\mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda\psi(\boldsymbol{\alpha}_i),$$

but the matrix factorization point of view allows us to make connections with numerous other unsupervised learning techniques, such as K-means, PCA, NMF, ICA...

Dictionary learning

Constrained variants

The formulations below are not equivalent

$$\min_{\mathbf{D}\in\mathbf{C},\mathbf{A}\in\mathbb{R}^{p\times n}} \sum_{i=1}^{n} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 \;\; \text{s.t.} \;\; \psi(\boldsymbol{\alpha}_i) \leq \mu.$$

or

$$\min_{\mathbf{D}\in\mathbf{C},\mathbf{A}\in\mathbb{R}^{p\times n}} \sum_{i=1}^{n} \psi(\boldsymbol{\alpha}_i) \;\; \text{s.t.} \;\; \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 \leq \varepsilon.$$

Using one instead of another depends on **the problem at hand**.

Pre-processing of natural image patches

Centering (also called removing the DC component)

$$\mathbf{x}_i \leftarrow \mathbf{x}_i - \left(\frac{1}{m} \sum_{j=1}^{m} \mathbf{x}_i[j] \right) \mathbf{1}_m,$$

(a) Without pre-processing.

(b) After centering.

Pre-processing of natural image patches

Contrast (variance) normalization

$$\mathbf{x}_i \leftarrow \frac{1}{\max(\|\mathbf{x}_i\|_2, \eta)} \mathbf{x}_i.$$

ex: η can be 0.2 times the mean value of the $\|\mathbf{x}_i\|_2$.

(a) After centering.

(b) After contrast normalization.

Pre-processing of natural image patches

Whitening after centering

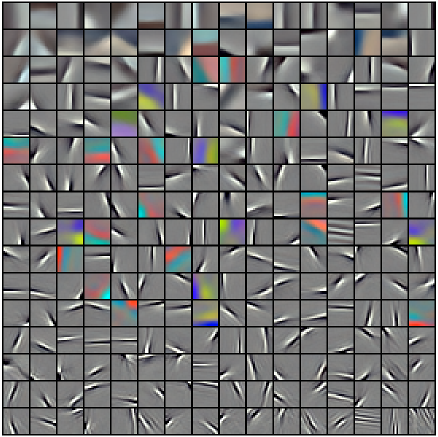
$$\mathbf{x}_i \leftarrow \mathbf{U}\mathbf{S}^{\dagger}\mathbf{U}^{\top}\mathbf{x}_i,$$

where $(1/\sqrt{n})\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\top}$ (SVD). Sometimes, small singular values are also set to zero. The resulting covariance $(1/n)\mathbf{X}\mathbf{X}^{\top}$ is close to identity.

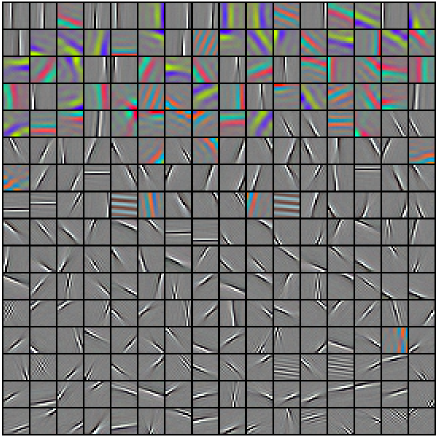
(a) After centering.

(b) After whitening.

Dictionary learning on color image patches



(c) With centering - RGB.

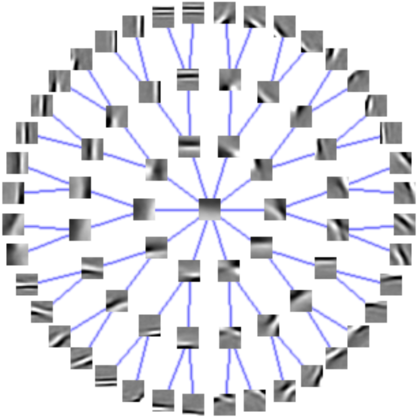


(d) With whitening - RGB.

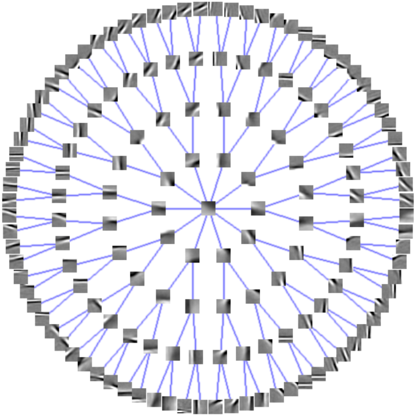
Figure: Dictionaries learned on RGB patches.

Dictionary learning with structured sparsity

Hierarchical dictionary learning



(a) Tree structure 1.

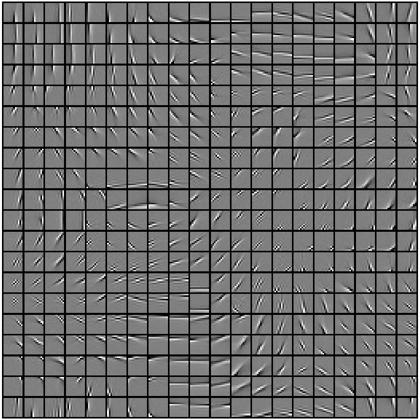


(b) Tree structure 2.

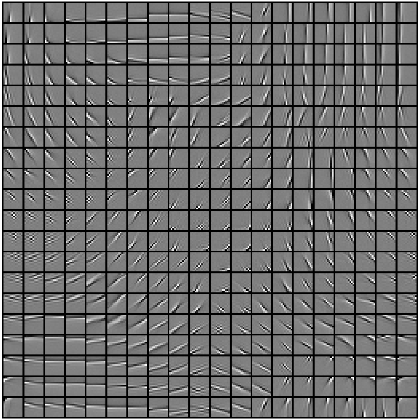
Figure: Hierarchical dictionaries learned on natural image patches of size 16×16 pixels.

Dictionary learning with structured sparsity

Topographic dictionary learning



(a) With 3×3 neighborhoods.



(b) With 4×4 neighborhood.

Figure: Topographic dictionaries learned on whitened natural image patches of size 12×12 pixels.

Part III: Sparse models for image processing

Image denoising

$$\underbrace{\mathbf{y}}_{\text{measurements}} = \underbrace{\mathbf{x}_{orig}}_{\text{original image}} + \underbrace{\mathbf{w}}_{\text{noise}}$$

Image denoising

Classical image models

$$\underbrace{\mathbf{y}}_{\text{measurements}} = \underbrace{\mathbf{x}_{orig}}_{\text{original image}} + \underbrace{\mathbf{w}}_{\text{noise}}.$$

Energy minimization problem - MAP estimation

$$E(\mathbf{x}) = \underbrace{\frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2}_{\text{relation to measurements}} + \underbrace{\psi(\mathbf{x})}_{\text{image model}}.$$

Some classical priors

- Smoothness $\lambda\|\mathcal{L}\mathbf{x}\|_2^2$;
- total variation $\lambda\|\nabla\mathbf{x}\|_1^2$ [Rudin et al., 1992];
- Markov random fields [Zhu and Mumford, 1997];
- wavelet sparsity $\lambda\|\mathbf{W}\mathbf{x}\|_1$.

Image denoising

The method of Elad and Aharon [2006]

Given a **fixed** dictionary $\mathbf{D}$, a patch $\mathbf{y}_i$ is denoised as follows:

1. center $\mathbf{y}_i$,
$$\mathbf{y}_i^c \triangleq \mathbf{y}_i - \mu_i \mathbf{1}_m \quad \text{with} \quad \mu_i \triangleq \frac{1}{n}\mathbf{1}_m^\top \mathbf{y}_i;$$
2. find a sparse linear combination of dictionary elements that approximates $\mathbf{y}_i^c$ up to the noise level:
$$\min_{\boldsymbol{\alpha}_i \in \mathbb{R}^p} \|\boldsymbol{\alpha}_i\|_0 \ \text{ s.t. } \ \|\mathbf{y}_i^c - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 \leq \varepsilon, \tag{1}$$
where ε is proportional to the noise variance σ^2;
3. add back the mean component to obtain the clean estimate $\hat{\mathbf{x}}_i$:
$$\hat{\mathbf{x}}_i \triangleq \mathbf{D}\boldsymbol{\alpha}_i^\star + \mu_i \mathbf{1}_m,$$

Image denoising

The method of Elad and Aharon [2006]

An **adaptive** approach

1. extract all overlapping $\sqrt{m} \times \sqrt{m}$ patches $\mathbf{y}_i$.
2. **dictionary learning**: learn $\mathbf{D}$ on the set of centered noisy patches $[\mathbf{y}_1^c, \ldots, \mathbf{y}_n^c]$.
3. **final reconstruction**: find an estimate $\hat{\mathbf{x}}_i$ for every patch using the approach of the previous slide;
4. **patch averaging**:

$$\hat{\mathbf{x}} = \frac{1}{m} \sum_{i=1}^{n} \mathbf{R}_i^\top \hat{\mathbf{x}}_i,$$

Remark

Like other state-of-the-art denoising approaches, it is patch-based [Buades et al., 2005, Dabov et al., 2007].

Practical tricks

- use larger patches when the noise level is high;
- choose $\varepsilon = m(1.15\sigma)^2$ or take the 0.9-quantile of the χ^2_m-distribution.
- always use the ℓ_0 regularization for the final reconstruction;
- using ℓ_1 for learning the dictionary seems to yield better results.

Image inpainting

[Mairal et al., 2008a,b]

For removing small holes in the image, a natural extension consists in introducing a **binary mask** $\mathbf{M}_i$ in the formulation:

$$\min_{\mathbf{D}\in\mathcal{C},\mathbf{A}\in\mathbb{R}^{p\times n}} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\|\mathbf{M}_i(\mathbf{y}_i-\mathbf{D}\boldsymbol{\alpha}_i)\|_2^2+\lambda\psi(\boldsymbol{\alpha}_i),$$

The approach assumes that

- the noise is not structured;
- the holes are smaller than the patch size.

The problem is called inpainting [Bertalmio et al., 2000].

Image inpainting

[Mairal et al., 2008a,b]

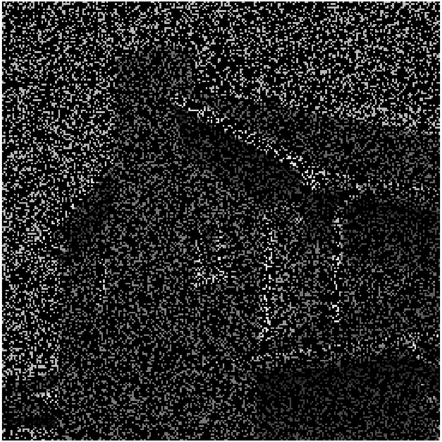


Image inpainting

[Mairal et al., 2008a,b]

Image inpainting

[Mairal et al., 2008a,b]

Image inpainting

Inpainting a 12-Mpixel photograph [Mairal et al., 2009a]

Image inpainting

Inpainting a 12-Mpixel photograph [Mairal et al., 2009a]

Image inpainting

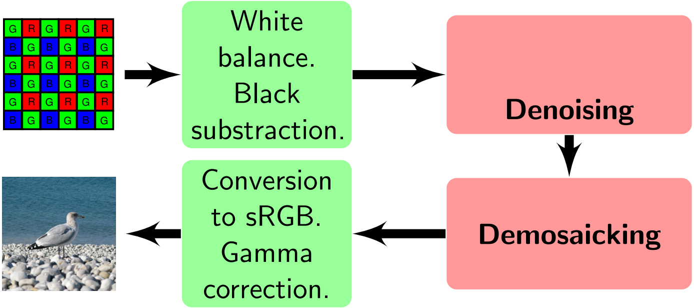
Inpainting a 12-Mpixel photograph [Mairal et al., 2009a]

Image inpainting

Inpainting a 12-Mpixel photograph [Mairal et al., 2009a]

Image demoisaicking

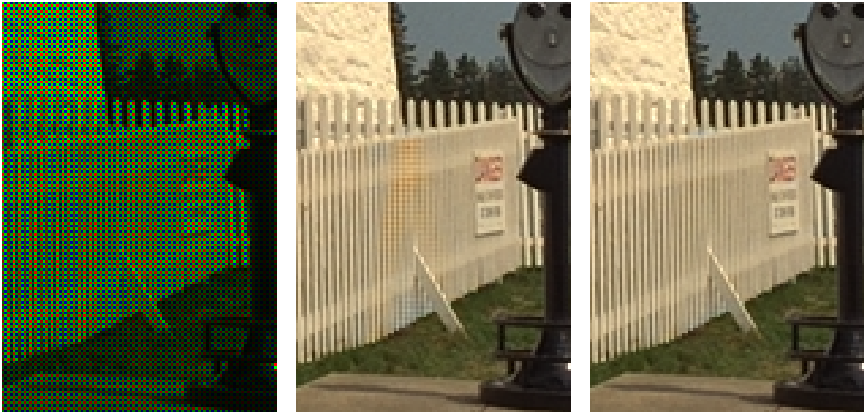
RAW Image Processing



Problem

The noise pattern is very structured: the previous inpainting scheme needs to be modified [Mairal et al., 2008a].

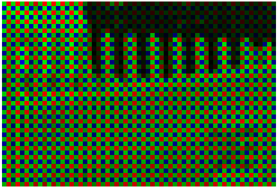
Image demoisaicking



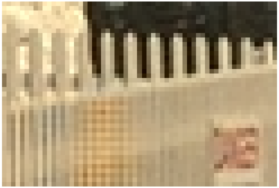
(a) Mosaicked image (b) Demosaicked image A (c) Demosaicked image B

Figure: Demosaicked image A is with the approach previously described; image B is with an extension called non-local sparse model [Mairal et al., 2009b].

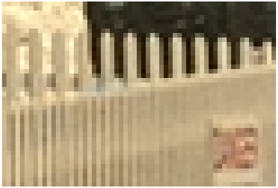
Image demoisaicking



(a) Zoom



(b) Zoom



(c) Zoom

Figure: Demosaicked image A is with the approach previously described; image B is with an extension called non-local sparse model [Mairal et al., 2009b].

Video processing

Extension developed by Protter and Elad [2009]:

Key ideas for video processing

- Using a 3D dictionary.
- Processing of many frames at the same time.
- Dictionary propagation.

Video processing

Inpainting, [Mairal et al., 2008b]

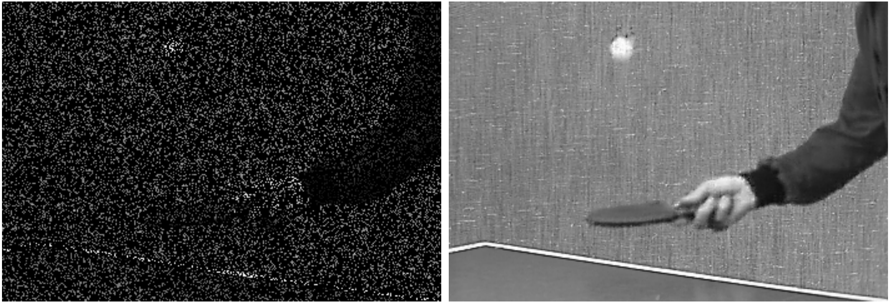


Figure: Inpainting results.

Video processing

Inpainting, [Mairal et al., 2008b]

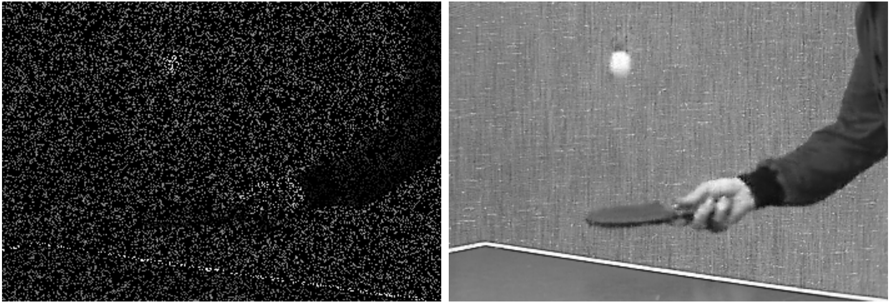
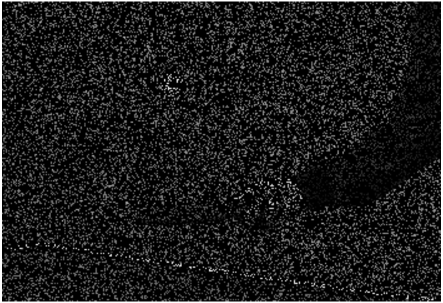


Figure: Inpainting results.

Video processing

Inpainting, [Mairal et al., 2008b]



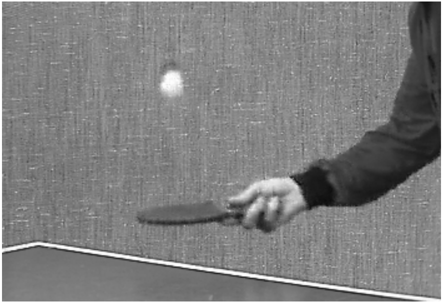


Figure: Inpainting results.

Video processing

Inpainting, [Mairal et al., 2008b]

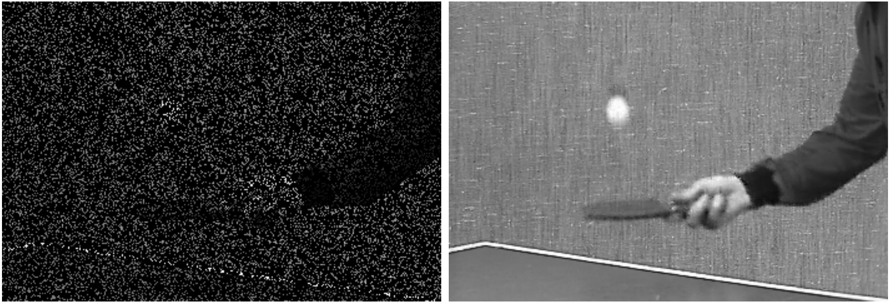
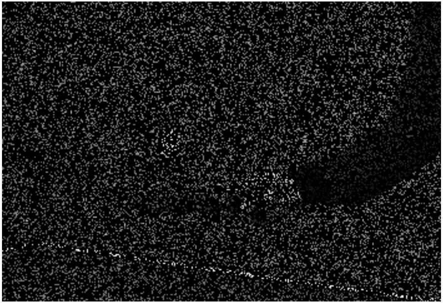


Figure: Inpainting results.

Video processing

Inpainting, [Mairal et al., 2008b]



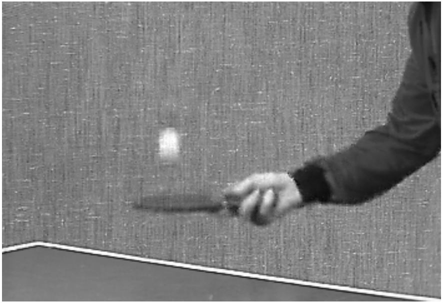


Figure: Inpainting results.

Video processing

Color video denoising, [Mairal et al., 2008b]

Figure: Inpainting results.

Video processing

Color video denoising, [Mairal et al., 2008b]

Figure: Inpainting results.

Video processing

Color video denoising, [Mairal et al., 2008b]

Figure: Inpainting results.

Video processing

Color video denoising, [Mairal et al., 2008b]

Figure: Inpainting results.

Video processing

Color video denoising, [Mairal et al., 2008b]

Figure: Inpainting results.

Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

Figure: Original

Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

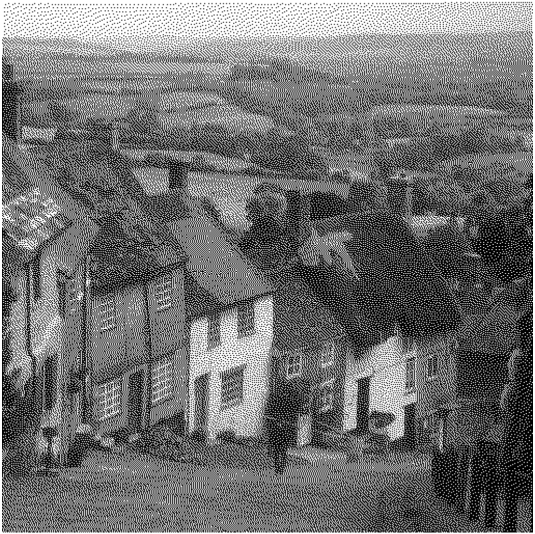


Figure: Binary image

Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

Figure: Reconstructed.

Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

Figure: Original

Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

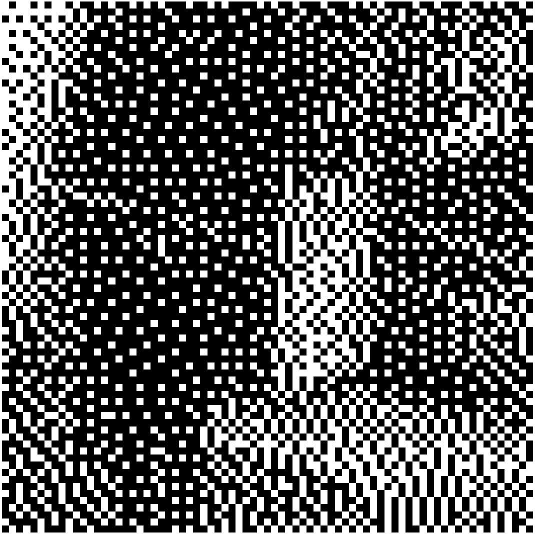


Figure: Binary image

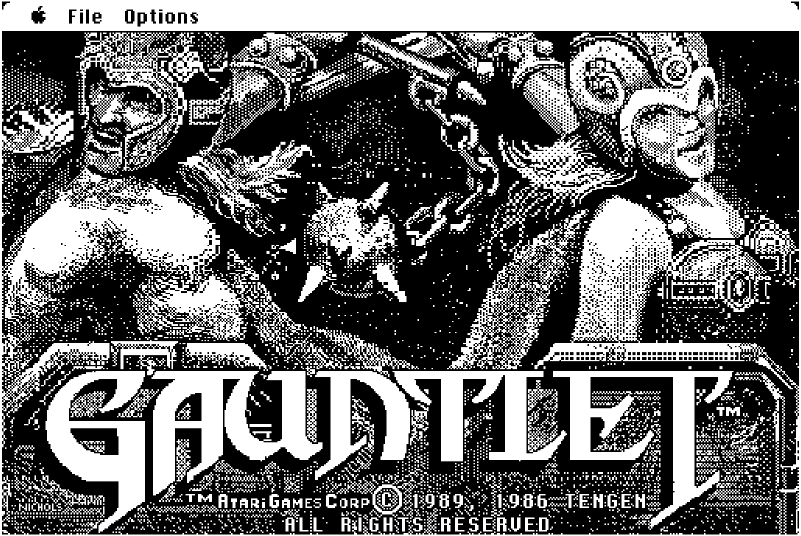
Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

Figure: Reconstructed.

Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]



Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

Inverting nonlinear local transformations

Inverse half-toning [Mairal et al., 2012]

Other patch modeling approaches

Non-local means and non-parametric approaches

Image pixels are well explained by a Nadaraya-Watson estimator:

$$\hat{\mathbf{x}}[i] = \sum_{j=1}^{n} \frac{K_h(\mathbf{y}_i - \mathbf{y}_j)}{\sum_{l=1}^{n} K_h(\mathbf{y}_i - \mathbf{y}_l)} \mathbf{y}[j], \tag{2}$$

with successful application to

- texture synthesis: [Efros and Leung, 1999]
- image denoising **(Non-local means)**: [Buades et al., 2005]
- image demosaicking: [Buades et al., 2009].

Other patch modeling approaches

BM3D

state-of-the-art image denoising approach [Dabov et al., 2007]:

- **block matching**: for each patch, find similar ones in the image;
- **3D wavelet filtering**: denoise blocks of patches with 3D-DCT;
- **patch averaging**: average estimates of overlapping patches;
- **second step with Wiener filtering**: use the first estimate to perform again and improve the previous steps.

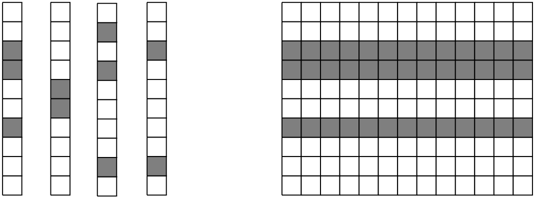
Further refined by Dabov et al. [2009] with shape-adaptive patches and PCA filtering.

Other patch modeling approaches

Non-local sparse models [Mairal et al., 2009b]

Exploit some ideas of BM3D to combine the non-local means principle with dictionary learning.

The main idea is that **similar patches should admit similar decompositions** by using group sparsity:



The approach uses a block matching/clustering step, followed by group sparse coding and patch averaging.

Other patch modeling approaches

Non-local sparse image models

Other patch modeling approaches

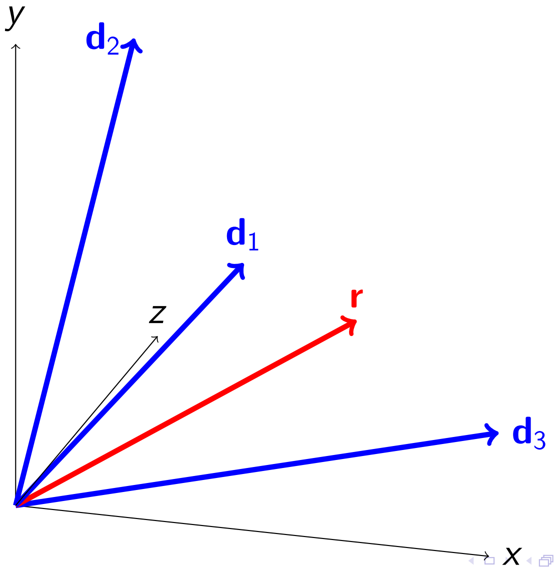
Non-local sparse image models

Part IV: Optimization for sparse estimation

Sparse reconstruction with the ℓ_0-penalty

Matching pursuit [Mallat and Zhang, 1993]

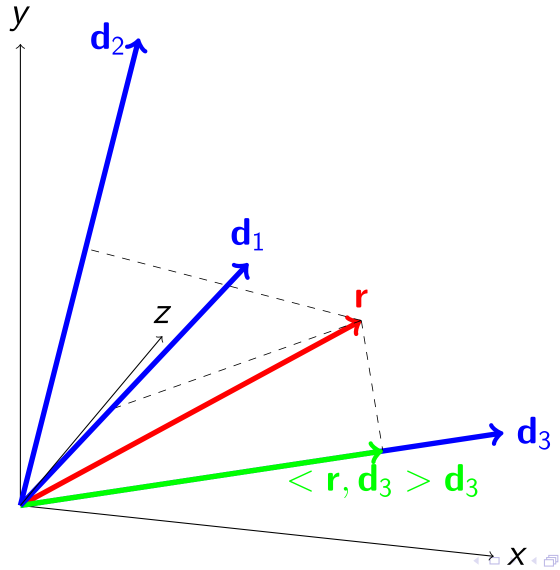
$$\boldsymbol{\alpha} = (0, 0, 0)$$



Sparse reconstruction with the ℓ_0-penalty

Matching pursuit [Mallat and Zhang, 1993]

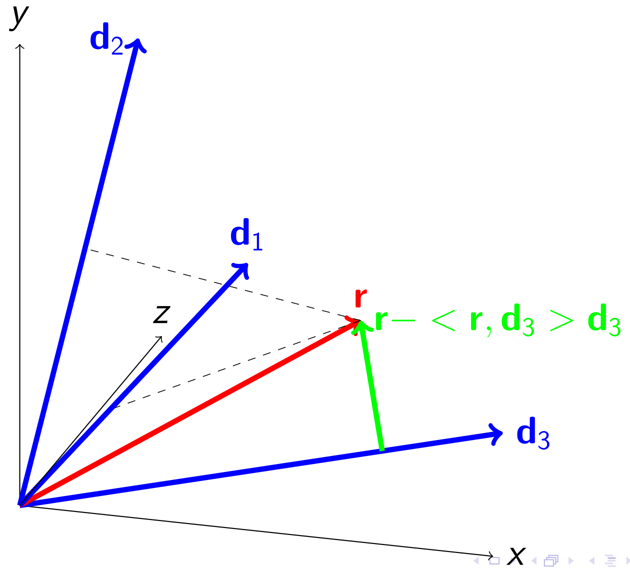
$$\boldsymbol{\alpha} = (0, 0, 0)$$



Sparse reconstruction with the ℓ_0-penalty

Matching pursuit [Mallat and Zhang, 1993]

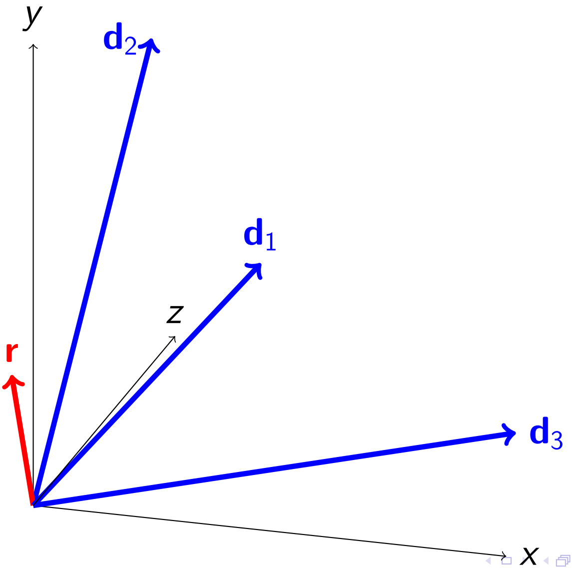
$$\boldsymbol{\alpha} = (0, 0, 0)$$



Sparse reconstruction with the ℓ_0-penalty

Matching pursuit [Mallat and Zhang, 1993]

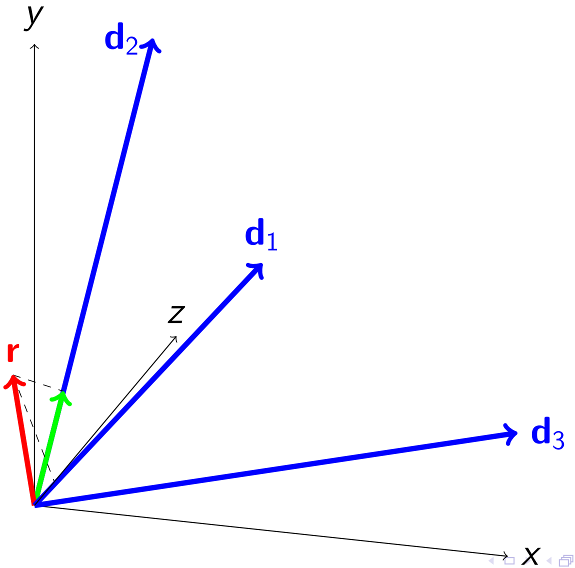
$$\boldsymbol{\alpha} = (0, 0, 0.75)$$



Sparse reconstruction with the ℓ_0-penalty

Matching pursuit [Mallat and Zhang, 1993]

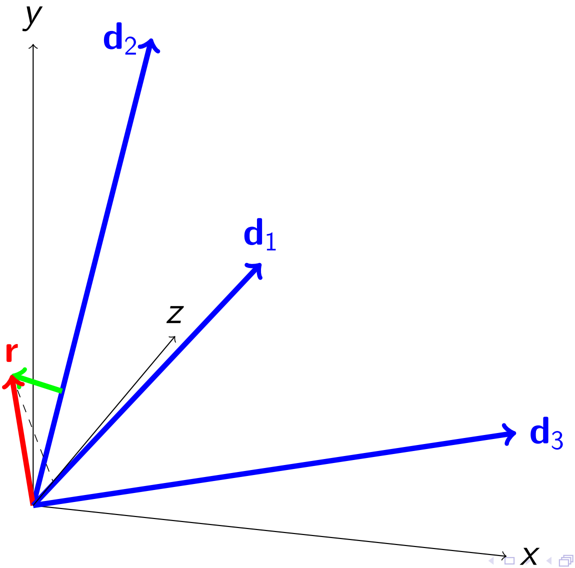
$$\boldsymbol{\alpha} = (0, 0, 0.75)$$



Sparse reconstruction with the ℓ_0-penalty

Matching pursuit [Mallat and Zhang, 1993]

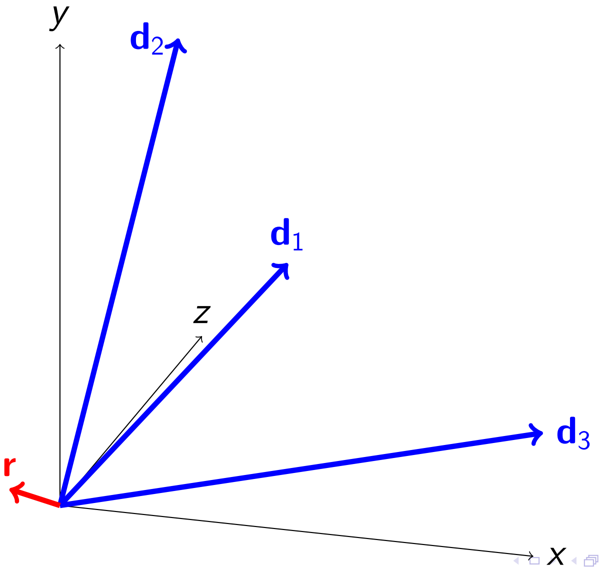
$$\boldsymbol{\alpha} = (0, 0, 0.75)$$



Sparse reconstruction with the ℓ_0-penalty

Matching pursuit [Mallat and Zhang, 1993]

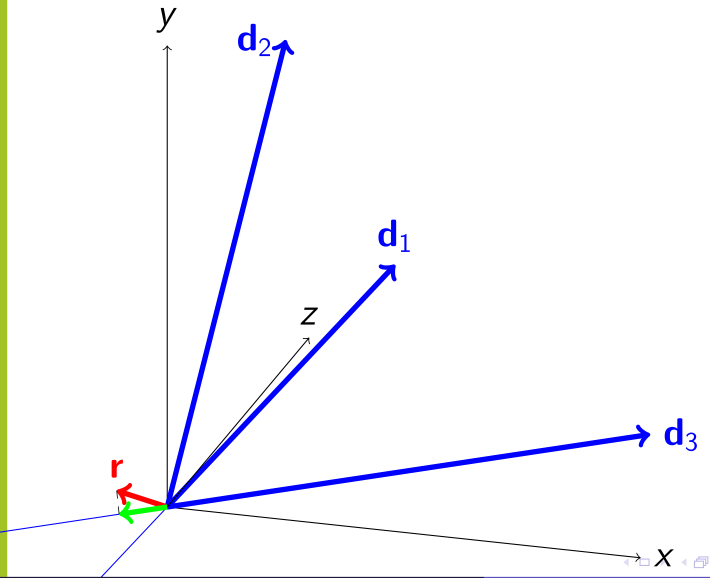
$$\boldsymbol{\alpha} = (0, 0.24, 0.75)$$



Sparse reconstruction with the ℓ_0-penalty

Matching pursuit [Mallat and Zhang, 1993]

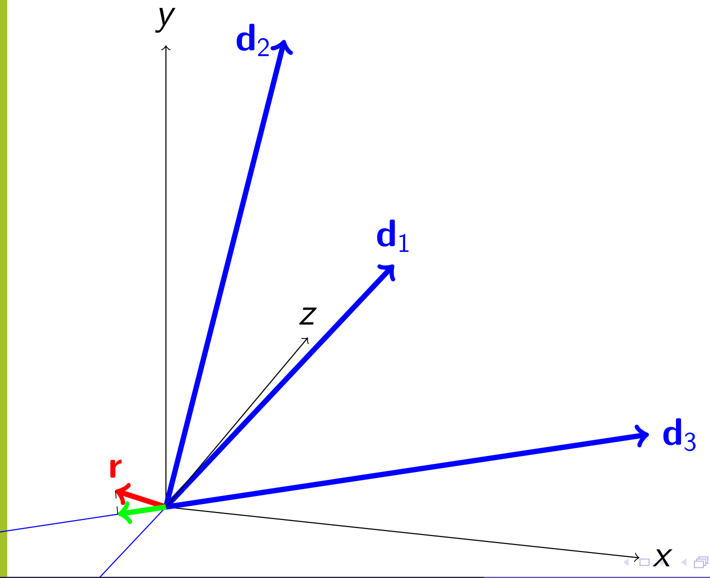
$$\boldsymbol{\alpha} = (0, 0.24, 0.75)$$



Sparse reconstruction with the ℓ_0-penalty

Matching pursuit [Mallat and Zhang, 1993]

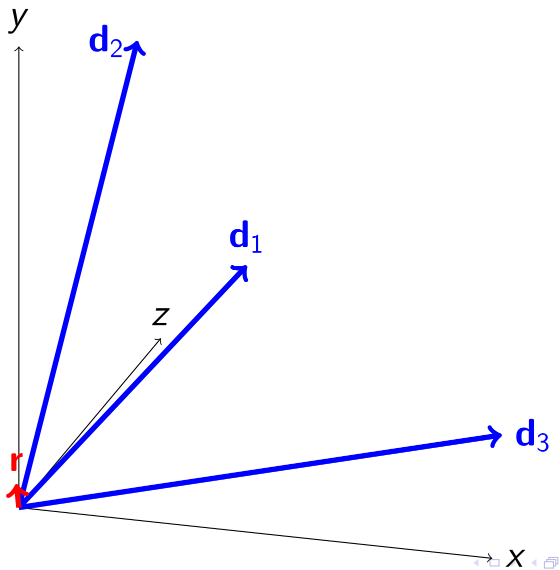
$$\boldsymbol{\alpha} = (0, 0.24, 0.75)$$



Sparse reconstruction with the ℓ_0-penalty

Matching pursuit [Mallat and Zhang, 1993]

$$\boldsymbol{\alpha} = (0, 0.24, 0.65)$$



Sparse reconstruction with the ℓ_0-penalty

Matching pursuit [Mallat and Zhang, 1993]

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\underbrace{\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}}_{\mathbf{r}}\|_2^2 \;\text{ s.t. }\; \|\boldsymbol{\alpha}\|_0 \leq k.$$

1: $\boldsymbol{\alpha} \leftarrow 0$
2: $\mathbf{r} \leftarrow \mathbf{x}$ (residual).
3: **while** $\|\boldsymbol{\alpha}\|_0 < k$ **do**
4: Select the predictor with maximum inner-product with the residual

$$\hat{\jmath} \leftarrow \arg\max_{j=1,\ldots,p} |\mathbf{d}_j^\top \mathbf{r}|$$

5: Update the residual and the coefficients

$$\begin{aligned} \boldsymbol{\alpha}[\hat{\jmath}] &\leftarrow \boldsymbol{\alpha}[\hat{\jmath}] + \mathbf{d}_{\hat{\jmath}}^\top \mathbf{r} \\ \mathbf{r} &\leftarrow \mathbf{r} - (\mathbf{d}_{\hat{\jmath}}^\top \mathbf{r})\mathbf{d}_{\hat{\jmath}} \end{aligned}$$

6: **end while**

Sparse reconstruction with the ℓ_0-penalty

Matching pursuit [Mallat and Zhang, 1993]

Remarks

- Matching pursuit is a **coordinate descent** algorithm. It greedily selects one coordinate at a time and optimizes the cost function with respect to that coordinate.

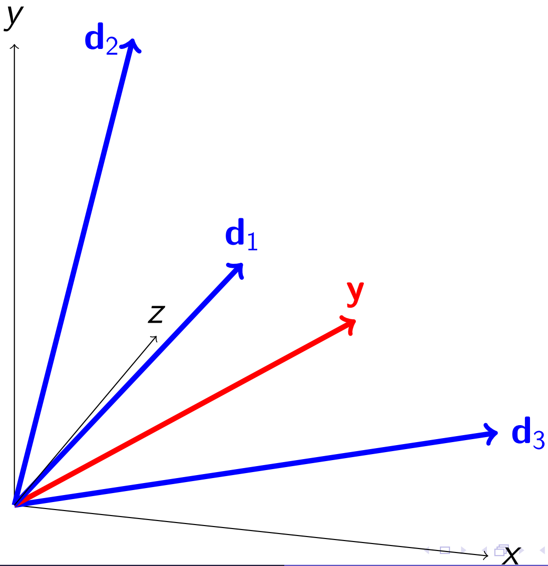
$$\boldsymbol{\alpha}[\hat{\jmath}] \leftarrow \operatorname*{arg\,min}_{\alpha \in \mathbb{R}} \left\| \mathbf{x} - \sum_{l \neq \hat{\jmath}} \boldsymbol{\alpha}[l]\mathbf{d}_l - \alpha \mathbf{d}_{\hat{\jmath}} \right\|_2^2.$$

- Each coordinate can be selected several times during the process.
- The roots of this algorithm can be found in the statistics literature [Efroymson, 1960].

Sparse reconstruction with the ℓ_0-penalty

Orthogonal matching pursuit [Pati et al., 1993]

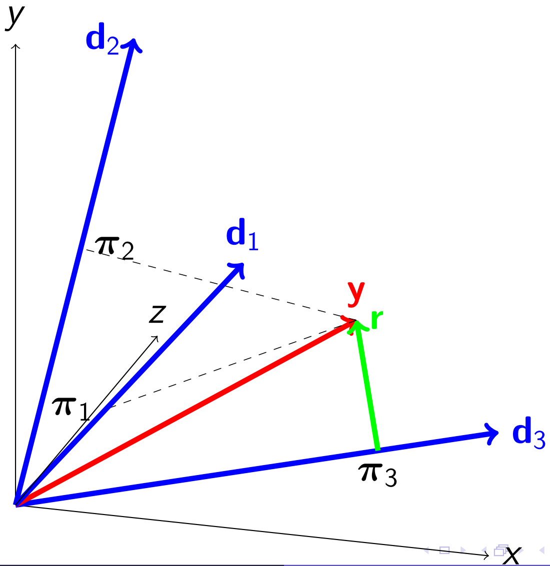
$$\boldsymbol{\alpha} = (0, 0, 0)$$
$$\Gamma = \emptyset$$



Sparse reconstruction with the ℓ_0-penalty

Orthogonal matching pursuit [Pati et al., 1993]

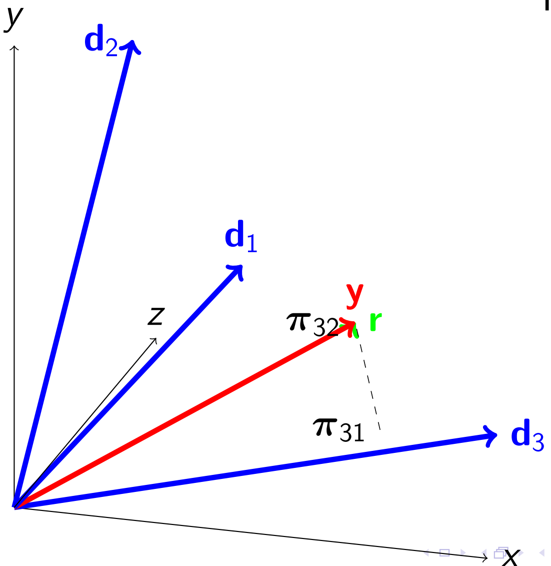
$$\boldsymbol{\alpha} = (0, 0, 0.75)$$
$$\Gamma = \{3\}$$



Sparse reconstruction with the ℓ_0-penalty

Orthogonal matching pursuit [Pati et al., 1993]

$$\boldsymbol{\alpha} = (0, 0.29, 0.63)$$
$$\Gamma = \{3, 2\}$$



Sparse reconstruction with the ℓ_0-penalty

Orthogonal matching pursuit [Pati et al., 1993]

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \;\; \text{s.t.} \;\; \|\boldsymbol{\alpha}\|_0 \leq k$$

$\Gamma = \emptyset$.

for $iter = 1, \ldots, k$ **do**

  Select the variable that most reduces the objective

$$(\hat{\jmath}, \hat{\boldsymbol{\beta}}) \leftarrow \operatorname*{arg\,min}_{j \in \Gamma^{\complement}, \boldsymbol{\beta}} \|\mathbf{x} - \mathbf{D}_{\Gamma \cup \{j\}} \boldsymbol{\beta}\|_2^2.$$

  Update the active set: $\Gamma \leftarrow \Gamma \cup \{\hat{\jmath}\}$.

  Update the coefficients:

$$\boldsymbol{\alpha}[\Gamma] \leftarrow \boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{\alpha}[\Gamma^{\complement}] \leftarrow 0.$$

end for

Sparse reconstruction with the ℓ_0-penalty

Orthogonal matching pursuit [Pati et al., 1993]

Remarks

- this is an **active-set** algorithm.
- when a new variable is selected, the coefficients for the full set Γ are re-optimized:

$$\boldsymbol{\alpha}[\Gamma] = (\mathbf{D}_\Gamma^\top \mathbf{D}_\Gamma)^{-1}\mathbf{D}_\Gamma^\top \mathbf{x},$$

  and the residual is always orthogonal to the matrix $\mathbf{D}_\Gamma$ of previously selected dictionary elements:

$$\mathbf{D}_\Gamma^\top(\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}) = \mathbf{D}_\Gamma^\top(\mathbf{x} - \mathbf{D}_\Gamma\boldsymbol{\alpha}[\Gamma]) = 0.$$

- several variants of OMP exist regarding the selection rule of $\hat{\jmath}$. The one we use appears in Cotter et al. [1999].

Sparse reconstruction with the ℓ_0-penalty

Orthogonal matching pursuit [Pati et al., 1993]

Keys for a fast implementation

- If available, use the Gram matrix $\mathbf{G} = \mathbf{D}^\top \mathbf{D}$;
- Maintain the computation of $\mathbf{D}^\top(\mathbf{x} - \mathbf{D}\boldsymbol{\alpha})$,
- Update the Cholesky decomposition of $(\mathbf{D}_\Gamma^\top \mathbf{D}_\Gamma)^{-1}$.

The total complexity for decomposing n k-sparse signals of size m with a dictionary of size p is

$$\underbrace{O(p^2 m)}_{\text{Gram matrix}} + \underbrace{O(nk^3)}_{\text{Cholesky}} + \underbrace{O(n(pm + pk^2))}_{\mathbf{D}^\top(\mathbf{x}-\mathbf{D}\boldsymbol{\alpha})} = O(np(m + k^2))$$

It is also possible to use the matrix inversion lemma instead of a Cholesky decomposition.

Sparse reconstruction with the ℓ_0-penalty

Orthogonal matching pursuit [Pati et al., 1993]

Example with the software SPAMS

Software available at `http://spams-devel.gforge.inria.fr/`.

```
>> I=double(imread('data/lena.eps'))/255;
>> %extract all patches of I
>> X=im2col(I,[8 8],'sliding');
>> %load a dictionary of size 64 x 256
>> D=load('dict.mat');
>>
>> %set the sparsity parameter L to 10
>> param.L=10;
>> alpha=mexOMP(X,D,param);
```

On this dual-core laptop: **150000 signals processed per second!**

Sparse reconstruction with the ℓ_1-norm

Coordinate descent for the Lasso [Fu, 1998]

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1.$$

The coordinate descent method consists of iteratively fixing all variables and optimizing with respect to one:

$$\boldsymbol{\alpha}[j] \leftarrow \arg\min_{\alpha \in \mathbb{R}} \frac{1}{2}\|\underbrace{\mathbf{x} - \sum_{l \neq j} \boldsymbol{\alpha}[l]\mathbf{d}_l}_{\mathbf{r}} - \alpha\mathbf{d}_j\|_2^2 + \lambda|\alpha|.$$

Assume the columns of $\mathbf{D}$ to have unit ℓ_2-norm,

$$\boldsymbol{\alpha}_j \leftarrow \operatorname{sign}(\mathbf{d}_j^\top \mathbf{r})(|\mathbf{d}_j^\top \mathbf{r}| - \lambda)^+$$

This involves again the **soft-thresholding** operator.

Optimization for Dictionary Learning

$$\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^{p\times n}\\ \mathbf{D}\in\mathcal{C}}} \sum_{i=1}^{n} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda\psi(\boldsymbol{\alpha}_i)$$

$$\mathcal{C} \triangleq \{\mathbf{D} \in \mathbb{R}^{m\times p} \ \text{ s.t. } \ \forall j = 1,\dots,p, \ \ \|\mathbf{d}_j\|_2 \leq 1\}.$$

Classical approach

- Alternate minimization between $\mathbf{D}$ and $\boldsymbol{\alpha}$ (MOD with $\psi = \ell_0$ [Engan et al., 1999], K-SVD with $\psi = \ell_0$ [Aharon et al., 2006], [Lee et al., 2007] with $\psi = \ell_1$);
- good results, reliable, but can be slow when n is large!

Conclusion

What we have seen:

- why the ℓ_1-norm induce sparsity (part I);
- the classical dictionary learning formulations on natural image patches (part II);
- a few applications to image restoration (part III);
- a few algorithms (part IV).

Conclusion

What we have NOT seen:

- structured sparsity, theory of sparse estimation.
- other matrix factorization formulations;
- applications in computer vision;
- many algorithms including stochastic optimization for dictionary learning.

References I

M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the ACM SIGGRAPH Conference*, 2000.

A. Buades, B. Coll, and J.-M. Morel. A review of image denoising algorithms, with a new one. *SIAM Journal on Multiscale Modeling and Simulation*, 4(2):490–530, 2005.

A. Buades, B. Coll, J.-M. Morel, and C. Sbert. Self-similarity driven color demosaicking. *IEEE Transactions on Image Processing*, 18(6): 1192–1202, 2009.

M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust. Do we know what the early visual system does? *The Journal of Neuroscience*, 25(46): 10577–10597, 2005.

References II

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1999.

J. F. Claerbout and F. Muir. Robust modeling with erratic data. *Geophysics*, 38(5):826–844, 1973.

S. F. Cotter, J. Adler, B. Rao, and K. Kreutz-Delgado. Forward sequential algorithms for best basis selection. In *IEEE Proceedings of Vision Image and Signal Processing*, pages 235–244, 1999.

K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007.

K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. BM3D image denoising with shape-adaptive principal component analysis. In *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*, 2009.

References III

J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169, 1985.

A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 1999.

M. A. Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, 9(1):191–203, 1960.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

K. Engan, S. O. Aase, and J. H. Husoy. Method of optimal directions for frame design. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999.

References IV

W.J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195 (1):215–243, 1968.

H. Lee, A. Battle, R. Raina, and A.Y. Ng. Efficient sparse coding algorithms. *Advances in Neural Information Processing Systems (NIPS)*, 19:801, 2007.

J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, 2008a.

J. Mairal, G. Sapiro, and M. Elad. Learning multiscale sparse representations for image and video restoration. *SIAM Multiscale Modeling and Simulation*, 7(1):214–241, 2008b.

References V

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009a.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009b.

J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4): 791–804, 2012.

S. Mallat and Z. Zhang. Matching pursuit in a time-frequency dictionary. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.

References VI

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381: 607–609, 1996.

B. A. Olshausen and D. J. Field. How close are we to understanding V1? *Neural computation*, 17(8):1665–1699, 2005.

Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, 1993.

M. Protter and M. Elad. Image sequence denoising via sparse and redundant representations. *IEEE Transactions on Image Processing*, 18(1):27–35, 2009.

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4): 259–268, 1992.

References VII

H. L. Taylor, S. C. Banks, and J. F. McCoy. Deconvolution with the ℓ_1 norm. *Geophysics*, 44(1):39–52, 1979.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.

D. Wrinch and H. Jeffreys. XLII. On certain fundamental principles of scientific inquiry. *Philosophical Magazine Series 6*, 42(249):369–390, 1921.

S. C. Zhu and D. Mumford. Prior learning and gibbs reaction-diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 (11):1236–1250, 1997.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2): 301–320, 2005.