

Sparse Coding and Dictionary Learning for Image Analysis

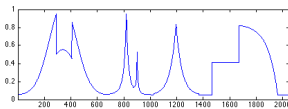
Part I: Optimization for Sparse Coding

Francis Bach, Julien Mairal, Jean Ponce and Guillermo Sapiro

ICCV'09 tutorial, Kyoto, 28th September 2009

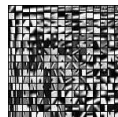
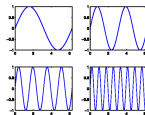
What is a Sparse Linear Model?

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



Let $\mathbf{D} = [\mathbf{d}_1, \dots, \mathbf{d}_p] \in \mathbb{R}^{m \times p}$ be a set of normalized “basis vectors”.

We call it **dictionary**.



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

$$\underbrace{\begin{pmatrix} \mathbf{x} \end{pmatrix}}_{\mathbf{x} \in \mathbb{R}^m} \approx \underbrace{\begin{pmatrix} \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_p \end{pmatrix}}_{\mathbf{D} \in \mathbb{R}^{m \times p}} \underbrace{\begin{pmatrix} \alpha[1] \\ \alpha[2] \\ \vdots \\ \alpha[p] \end{pmatrix}}_{\alpha \in \mathbb{R}^p, \text{ sparse}}$$

The Sparse Decomposition Problem

$$\min_{\alpha \in \mathbb{R}^p} \underbrace{\frac{1}{2} \|\mathbf{x} - \mathbf{D}\alpha\|_2^2}_{\text{data fitting term}} + \underbrace{\lambda \psi(\alpha)}_{\text{sparsity-inducing regularization}}$$

ψ induces sparsity in α . It can be

- the ℓ_0 “pseudo-norm”. $\|\alpha\|_0 \triangleq \#\{i \text{ s.t. } \alpha[i] \neq 0\}$ (NP-hard)
- the ℓ_1 norm. $\|\alpha\|_1 \triangleq \sum_{i=1}^p |\alpha[i]|$ (convex)
- ...

This is a **selection** problem.

Finding your way in the sparse coding literature. . .

. . . is not easy. The literature is vast, redundant, sometimes confusing and many papers are claiming victory. . .

The main class of methods are

- greedy procedures [Mallat and Zhang, 1993], [Weisberg, 1980]
- homotopy [Osborne et al., 2000], [Efron et al., 2004], [Markowitz, 1956]
- soft-thresholding based methods [Fu, 1998], [Daubechies et al., 2004], [Friedman et al., 2007], [Nesterov, 2007], [Beck and Teboulle, 2009], . . .
- reweighted- ℓ_2 methods [Daubechies et al., 2009], . . .
- active-set methods [Roth and Fischer, 2008].
- . . .

- 1 Greedy Algorithms
- 2 Homotopy and LARS
- 3 Soft-thresholding based optimization

Matching Pursuit

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

- 1: $\alpha \leftarrow 0$
- 2: $\mathbf{r} \leftarrow \mathbf{x}$ (residual).
- 3: **while** $\|\alpha\|_0 < L$ **do**
- 4: Select the atom with maximum correlation with the residual

$$\hat{i} \leftarrow \arg \max_{i=1,\dots,p} |\mathbf{d}_i^T \mathbf{r}|$$

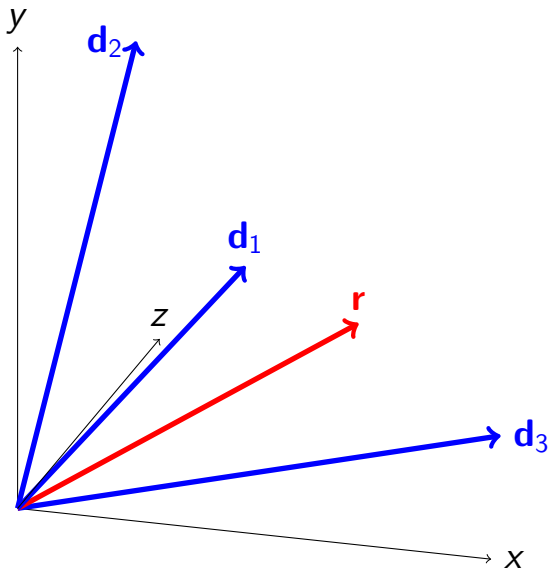
- 5: Update the residual and the coefficients

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

- 6: **end while**

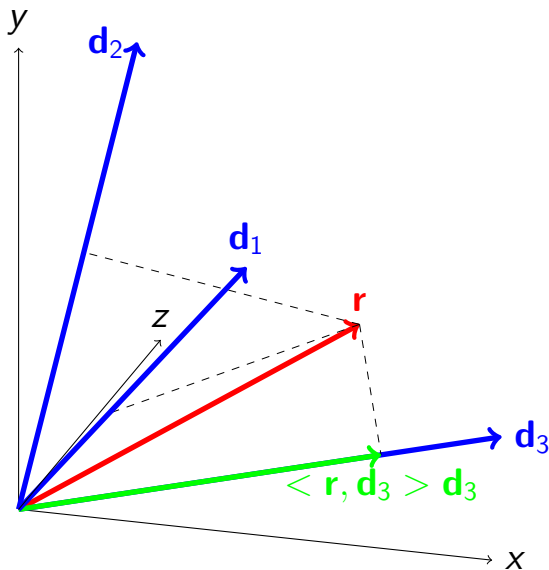
Matching Pursuit

$$\alpha = (0, 0, 0)$$



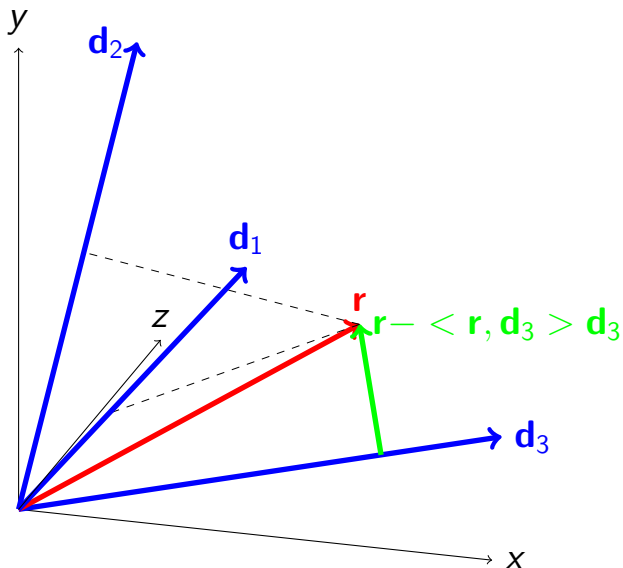
Matching Pursuit

$$\alpha = (0, 0, 0)$$



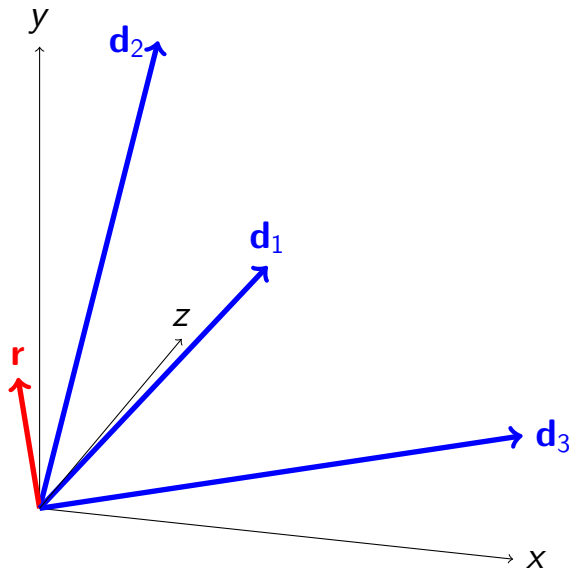
Matching Pursuit

$$\alpha = (0, 0, 0)$$



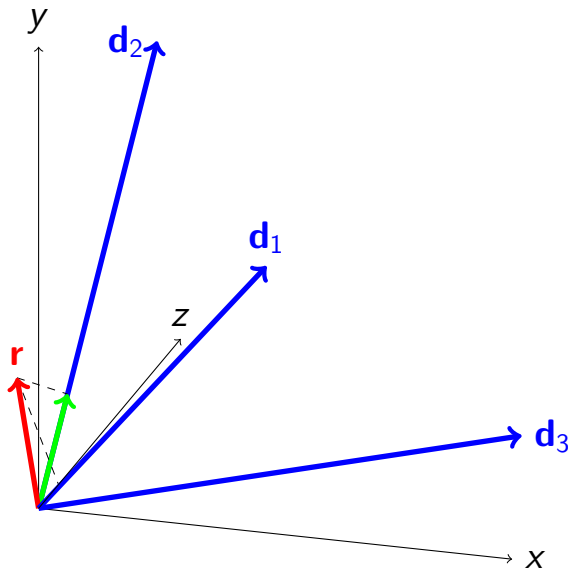
Matching Pursuit

$$\alpha = (0, 0, 0.75)$$



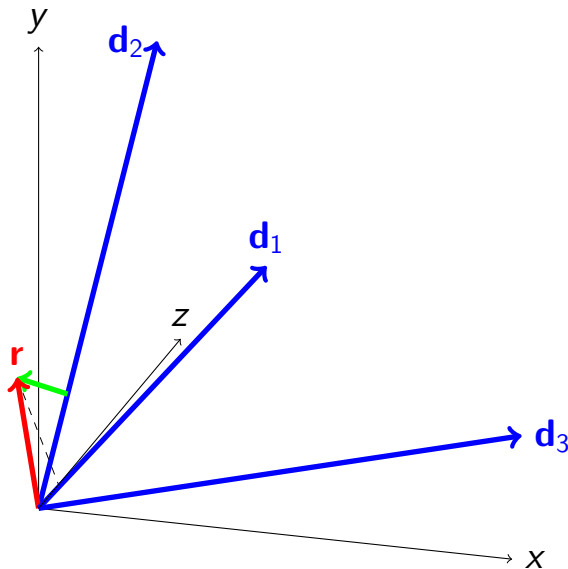
Matching Pursuit

$$\alpha = (0, 0, 0.75)$$



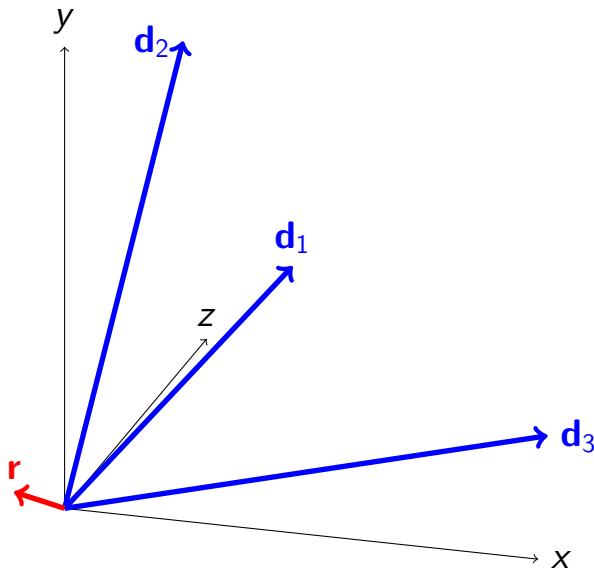
Matching Pursuit

$$\alpha = (0, 0, 0.75)$$



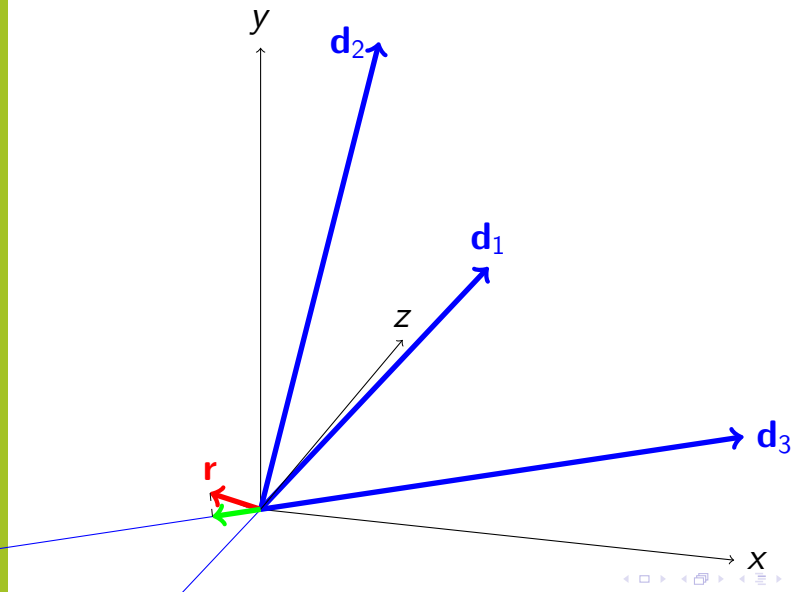
Matching Pursuit

$$\alpha = (0, 0.24, 0.75)$$



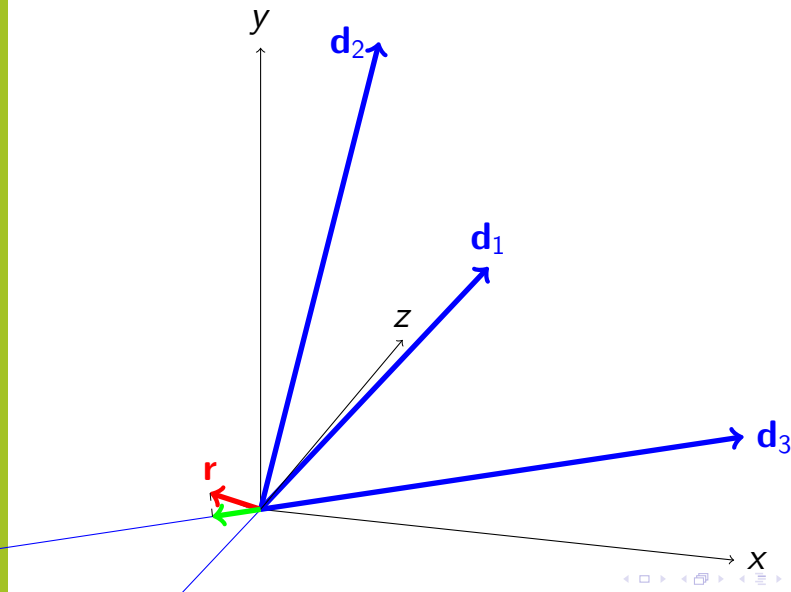
Matching Pursuit

$$\alpha = (0, 0.24, 0.75)$$



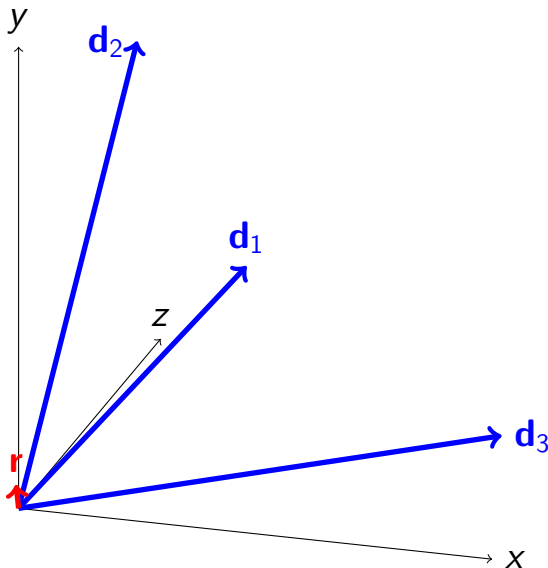
Matching Pursuit

$$\alpha = (0, 0.24, 0.75)$$



Matching Pursuit

$$\alpha = (0, 0.24, 0.65)$$



Orthogonal Matching Pursuit

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

- 1: $\Gamma = \emptyset$.
- 2: **for** $iter = 1, \dots, L$ **do**
- 3: Select the atom which most reduces the objective

$$\hat{i} \leftarrow \arg \min_{i \in \Gamma^c} \left\{ \min_{\alpha'} \|\mathbf{x} - \mathbf{D}_{\Gamma \cup \{i\}} \alpha'\|_2^2 \right\}$$

- 4: Update the active set: $\Gamma \leftarrow \Gamma \cup \{\hat{i}\}$.
- 5: Update the residual (orthogonal projection)

$$\mathbf{r} \leftarrow (\mathbf{I} - \mathbf{D}_\Gamma (\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1} \mathbf{D}_\Gamma^T) \mathbf{x}.$$

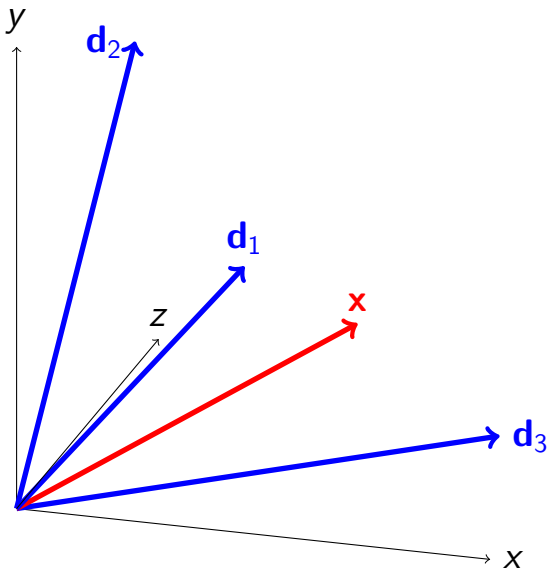
- 6: Update the coefficients

$$\alpha_\Gamma \leftarrow (\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1} \mathbf{D}_\Gamma^T \mathbf{x}.$$

- 7: **end for**

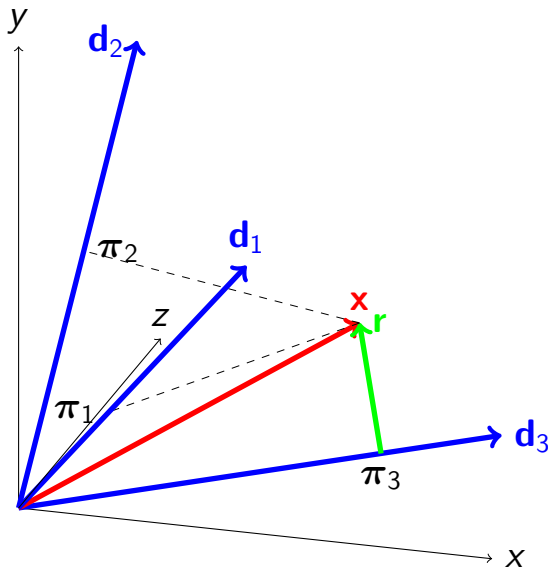
Orthogonal Matching Pursuit

$$\alpha = (0, 0, 0)$$
$$\Gamma = \emptyset$$



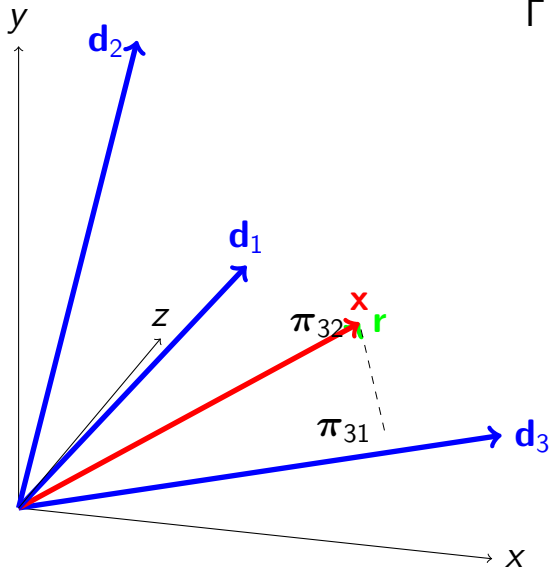
Orthogonal Matching Pursuit

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



Orthogonal Matching Pursuit

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



Orthogonal Matching Pursuit

Contrary to MP, an atom can only be selected one time with OMP. It is, however, more difficult to implement efficiently. The keys for a good implementation in the case of a large number of signals are

- Precompute the Gram matrix $\mathbf{G} = \mathbf{D}^T \mathbf{D}$ once in for all,
- Maintain the computation of $\mathbf{D}^T \mathbf{r}$ for each signal,
- Maintain a Cholesky decomposition of $(\mathbf{D}_r^T \mathbf{D}_r)^{-1}$ for each signal.

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

$$\underbrace{O(p^2 m)}_{\text{Gram matrix}} + \underbrace{O(nL^3)}_{\text{Cholesky}} + \underbrace{O(n(pm + pL^2))}_{\mathbf{D}^T \mathbf{r}} = O(np(m + L^2))$$

It is also possible to use the matrix inversion lemma instead of a Cholesky decomposition (same complexity, but less numerical stability)

Example with the software SPAMS

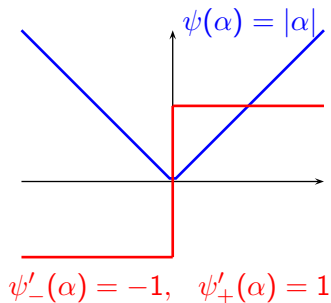
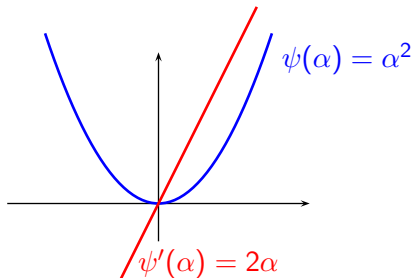
Software available at <http://www.di.ens.fr/willow/SPAMS/>

```
>> I=double(imread('data/lena.png'))/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 a 8-cores 2.83Ghz machine: **230000 signals processed per second!**

Why does the ℓ_1 -norm induce sparsity?

Analysis of the norms in 1D



The gradient of the ℓ_2 -norm 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?

Example: quadratic problem in 1D

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

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:

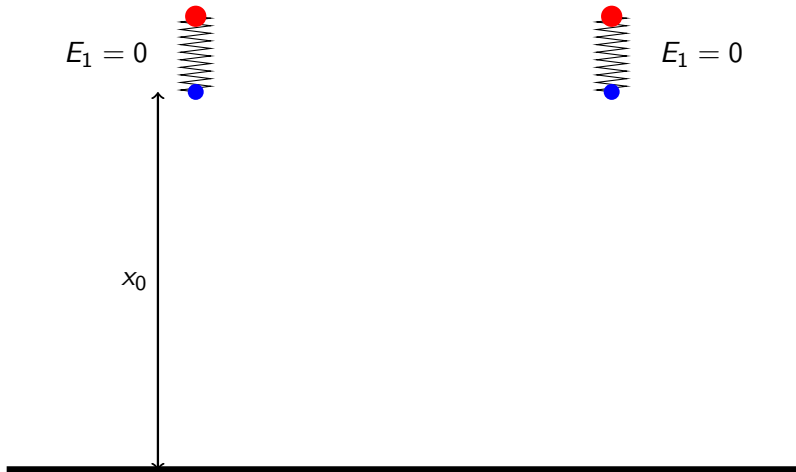
- $|\alpha| > 0$ and $(x - \alpha) + \lambda \operatorname{sign}(\alpha) = 0$
- $\alpha = 0$ and $g_+ \geq 0$ and $g_- \leq 0$

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

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

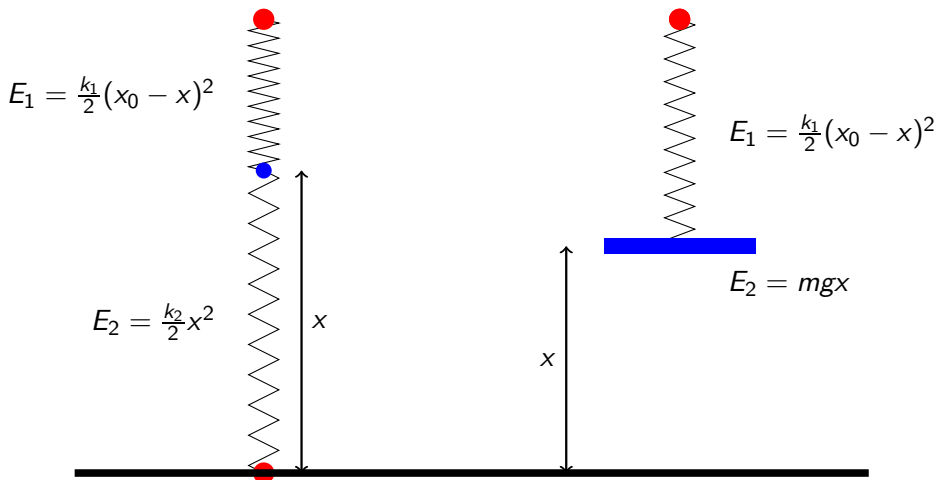
Why does the ℓ_1 -norm induce sparsity?

Physical illustration



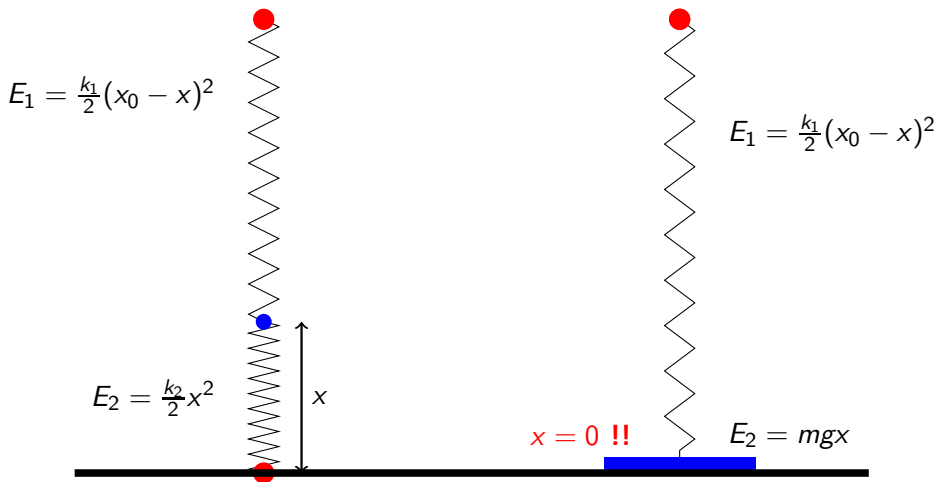
Why does the ℓ_1 -norm induce sparsity?

Physical illustration



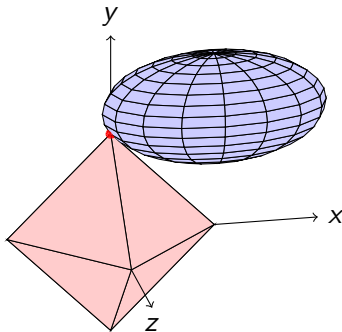
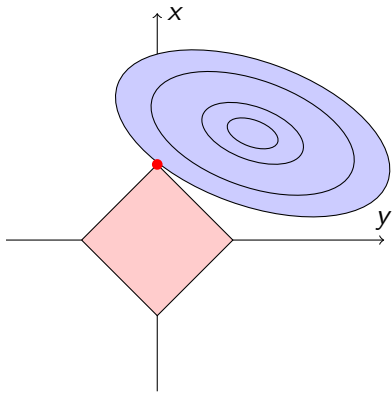
Why does the ℓ_1 -norm induce sparsity?

Physical illustration



Why does the ℓ_1 -norm induce sparsity?

The geometric explanation



general quadratic problem: **coupled** soft-thresholding.

Optimality conditions of the Lasso

Nonsmooth optimization

Directional derivatives and subgradients are useful tools for studying ℓ_1 -decomposition problems:

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

In this tutorial, we use the **directional derivatives** to derive simple optimality conditions of the Lasso.

For more information on convex analysis and nonsmooth optimization, see the following books: [Boyd and Vandenberghe, 2004], [Nocedal and Wright, 2006], [Borwein and Lewis, 2006], [Bonnans et al., 2006], [Bertsekas, 1999].

Optimality conditions of the Lasso

Directional derivatives

- **Directional derivative** in the direction \mathbf{u} at α :

$$\nabla f(\alpha, \mathbf{u}) = \lim_{t \rightarrow 0^+} \frac{f(\alpha + t\mathbf{u}) - f(\alpha)}{t}$$

- Main idea: in non smooth situations, one may need to look at all directions \mathbf{u} and not simply p independent ones!
- **Proposition 1:** if f is differentiable in α , $\nabla f(\alpha, \mathbf{u}) = \nabla f(\alpha)^T \mathbf{u}$.
- **Proposition 2:** α is optimal iff for all \mathbf{u} in \mathbb{R}^p , $\nabla f(\alpha, \mathbf{u}) \geq 0$.

Optimality conditions of the Lasso

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

α^* is optimal iff for all \mathbf{u} in \mathbb{R}^p , $\nabla f(\alpha, \mathbf{u}) \geq 0$ —that is,

$$-\mathbf{u}^T \mathbf{D}^T (\mathbf{x} - \mathbf{D}\alpha^*) + \lambda \sum_{i, \alpha^*[i] \neq 0} \text{sign}(\alpha^*[i]) \mathbf{u}[i] + \lambda \sum_{i, \alpha^*[i] = 0} |\mathbf{u}_i| \geq 0,$$

which is equivalent to the following conditions:

$$\forall i = 1, \dots, p, \quad \begin{cases} |\mathbf{d}_i^T (\mathbf{x} - \mathbf{D}\alpha^*)| \leq \lambda & \text{if } \alpha^*[i] = 0 \\ \mathbf{d}_i^T (\mathbf{x} - \mathbf{D}\alpha^*) = \lambda \text{sign}(\alpha^*[i]) & \text{if } \alpha^*[i] \neq 0 \end{cases}$$

Homotopy

- A homotopy method provides a set of solutions indexed by a parameter.
- The regularization path $(\lambda, \alpha^*(\lambda))$ for instance!!
- It can be useful when the path has some “nice” properties (piecewise linear, piecewise quadratic).
- LARS [Efron et al., 2004] starts from a trivial solution, and follows the regularization path of the Lasso, which is **piecewise linear**.

Homotopy, LARS

[Osborne et al., 2000], [Efron et al., 2004]

$$\forall i = 1, \dots, p, \quad \begin{cases} |\mathbf{d}_i^T(\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}^*)| \leq \lambda & \text{if } \alpha^*[i] = 0 \\ \mathbf{d}_i^T(\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}^*) = \lambda \operatorname{sign}(\alpha^*[i]) & \text{if } \alpha^*[i] \neq 0 \end{cases} \quad (1)$$

The regularization path is piecewise linear:

$$\mathbf{D}_\Gamma^T(\mathbf{x} - \mathbf{D}_\Gamma\boldsymbol{\alpha}_\Gamma^*) = \lambda \operatorname{sign}(\boldsymbol{\alpha}_\Gamma^*)$$

$$\boldsymbol{\alpha}_\Gamma^*(\lambda) = (\mathbf{D}_\Gamma^T \mathbf{D}_\Gamma)^{-1}(\mathbf{D}_\Gamma^T \mathbf{x} - \lambda \operatorname{sign}(\boldsymbol{\alpha}_\Gamma^*)) = \mathbf{A} + \lambda \mathbf{B}$$

A simple interpretation of LARS

- Start from the trivial solution ($\lambda = \|\mathbf{D}^T \mathbf{x}\|_\infty, \boldsymbol{\alpha}^*(\lambda) = 0$).
- Maintain the computations of $|\mathbf{d}_i^T(\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}^*(\lambda))|$ for all i .
- Maintain the computation of the current direction \mathbf{B} .
- Follow the path by reducing λ until the next kink.

Example with the software SPAMS

<http://www.di.ens.fr/willow/SPAMS/>

```
>> I=double(imread('data/lena.png'))/255;
>> %extract all patches of I
>> X=normalize(im2col(I,[8 8],'sliding'));
>> %load a dictionary of size 64 x 256
>> D=load('dict.mat');
>>
>> %set the sparsity parameter lambda to 0.15
>> param.lambda=0.15;
>> alpha=mexLasso(X,D,param);
```

On a 8-cores 2.83Ghz machine: **77000 signals processed per second!**
Note that it can also solve **constrained** version of the problem. The complexity is more or less the same as OMP and uses the same tricks (Cholesky decomposition).

Coordinate Descent

- Coordinate descent + nonsmooth objective: **WARNING: not convergent in general**
- Here, the problem is equivalent to a convex smooth optimization problem with **separable** constraints

$$\min_{\alpha_+, \alpha_-} \frac{1}{2} \|\mathbf{x} - \mathbf{D}_+ \alpha_+ + \mathbf{D}_- \alpha_-\|_2^2 + \lambda \alpha_+^T \mathbf{1} + \lambda \alpha_-^T \mathbf{1} \quad \text{s.t.} \quad \alpha_-, \alpha_+ \geq 0.$$

- For this **specific** problem, coordinate descent is **convergent**.
- Supposing $\|\mathbf{d}_i\|_2 = 1$, updating the coordinate i :

$$\begin{aligned} \alpha[i] &\leftarrow \arg \min_{\beta} \frac{1}{2} \left\| \mathbf{x} - \underbrace{\sum_{j \neq i} \alpha[j] \mathbf{d}_j}_{\mathbf{r}} - \beta \mathbf{d}_i \right\|_2^2 + \lambda |\beta| \\ &\leftarrow \text{sign}(\mathbf{d}_i^T \mathbf{r}) (|\mathbf{d}_i^T \mathbf{r}| - \lambda)^+ \end{aligned}$$

- \Rightarrow **soft-thresholding!**

Example with the software SPAMS

<http://www.di.ens.fr/willow/SPAMS/>

```
>> I=double(imread('data/lena.png'))/255;
>> %extract all patches of I
>> X=normalize(im2col(I,[8 8],'sliding'));
>> %load a dictionary of size 64 x 256
>> D=load('dict.mat');
>>
>> %set the sparsity parameter lambda to 0.15
>> param.lambda=0.15;
>> param.tol=1e-2;
>> param.itermax=200;
>> alpha=mexCD(X,D,param);
```

On a 8-cores 2.83Ghz machine: **93000 signals processed per second!**

first-order/proximal methods

$$\min_{\alpha \in \mathbb{R}^p} f(\alpha) + \lambda \psi(\alpha)$$

- f is strictly convex and differentiable with a Lipschitz gradient.
- Generalize the idea of gradient descent

$$\alpha_{k+1} \leftarrow \arg \min_{\alpha \in \mathbb{R}} f(\alpha_k) + \nabla f(\alpha_k)^T (\alpha - \alpha_k) + \frac{L}{2} \|\alpha - \alpha_k\|_2^2 + \lambda \psi(\alpha).$$

- There exists an accelerated scheme (gradient method with “extrapolation”) [Nesterov, 2007, 1983]
- Both are implemented in SPAMS.
- suited for large-scale experiments.

Summary of this part

- Greedy methods can address directly the NP-hard ℓ_0 -decomposition problem.
- ℓ_1 can be used as a convex relaxation for ℓ_0 .
- Homotopy methods can be extremely efficient for small or medium-sized problems, or when the solution is very sparse.
- Coordinate descent provides in general quickly a solution with a small/medium precision, but gets slower when there is a lot of correlation in the dictionary.
- First order methods are very attractive in the large scale setting.
- Other good alternatives exists, active-set, reweighted ℓ_2 methods, stochastic variants, variants of OMP,...

References I

- A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- D. P. Bertsekas. *Nonlinear programming*. Athena Scientific Belmont, Mass, 1999.
- J.F. Bonnans, J.C. Gilbert, C. Lemarechal, and C.A. Sagastizabal. *Numerical optimization: theoretical and practical aspects*. Springer-Verlag New York Inc, 2006.
- J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization: Theory and examples*. Springer, 2006.
- S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math*, 57: 1413–1457, 2004.
- I. Daubechies, R. DeVore, M. Fornasier, and S. Gunturk. Iteratively re-weighted least squares minimization for sparse recovery. *Commun. Pure Appl. Math*, 2009.
- B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of statistics*, 32(2):407–499, 2004.

References II

- J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of statistics*, 1(2):302–332, 2007.
- W. J. Fu. Penalized regressions: The bridge versus the Lasso. *Journal of computational and graphical statistics*, 7:397–416, 1998.
- S. Mallat and Z. Zhang. Matching pursuit in a time-frequency dictionary. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.
- H. M. Markowitz. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly*, 3:111–133, 1956.
- Y. Nesterov. Gradient methods for minimizing composite objective function. Technical report, CORE, 2007.
- Y. Nesterov. A method for solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Dokl.*, 27:372–376, 1983.
- J. Nocedal and SJ Wright. *Numerical Optimization*. Springer: New York, 2006. 2nd Edition.
- M. R. Osborne, B. Presnell, and B. A. Turlach. On the Lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–37, 2000.

References III

- V. Roth and B. Fischer. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.
- S. Weisberg. *Applied Linear Regression*. Wiley, New York, 1980.