Sparse Coding and Dictionary Learning for Image Analysis

Julien Mairal

INRIA Visual Recognition and Machine Learning Summer School, 27th July 2010
What this lecture is about?

- **Why sparsity, what for and how?**
- **Signal and image processing**: Restoration, reconstruction.
- **Machine learning**: Selecting relevant features.
- **Computer vision**: Modelling the local appearance of image patches.
- **Computer vision**: Recent (and intriguing) results in bags of words models.
- **Optimization**: Solving challenging problems.
The Image Denoising Problem

\[ \mathbf{y} = \mathbf{x}_{\text{orig}} + \mathbf{w} \]

- \( \mathbf{y} \): measurements
- \( \mathbf{x}_{\text{orig}} \): original image
- \( \mathbf{w} \): noise
Sparse representations for image restoration

\[
\mathbf{y} = \mathbf{x}_{\text{orig}} + \mathbf{w}
\]

measurements
original image
noise

Energy minimization problem - MAP estimation

\[
E(\mathbf{x}) = \frac{1}{2} \| \mathbf{y} - \mathbf{x} \|^2_2 + Pr(\mathbf{x})
\]

relation to measurements
image model (-log prior)

Some classical priors

- Smoothness \( \lambda \| \mathcal{L} \mathbf{x} \|^2_2 \)
- Total variation \( \lambda \| \nabla \mathbf{x} \|^2_1 \)
- MRF priors
- \( \ldots \)
What is a Sparse Linear Model?

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

Let $D = [d_1, \ldots, d_p] \in \mathbb{R}^{m \times p}$ be a set of normalized “basis vectors”. We call it dictionary.

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

Why Sparsity?

A dictionary can be good for representing a class of signals, but not for representing white Gaussian noise.
The Sparse Decomposition Problem

\[
\min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \left\| x - D\alpha \right\|_2^2 + \lambda \psi(\alpha)
\]

\(\psi\) induces sparsity in \(\alpha\). It can be

- the \(\ell_0\) “pseudo-norm” \(\| \alpha \|_0 \overset{\triangle}{=} \# \{i \text{ s.t. } \alpha[i] \neq 0 \} \) (NP-hard)
- the \(\ell_1\) norm \(\| \alpha \|_1 \overset{\triangle}{=} \sum_{i=1}^{p} |\alpha[i]|\) (convex),
- \ldots

This is a selection problem. When \(\psi\) is the \(\ell_1\)-norm, the problem is called Lasso [Tibshirani, 1996] or basis pursuit [Chen et al., 1999]
Sparse representations for image restoration

Designed dictionaries

[Haar, 1910], [Zweig, Morlet, Grossman ∼70s], [Meyer, Mallat, Daubechies, Coifman, Donoho, Candes ∼80s-today]... (see [Mallat, 1999])
Wavelets, Curvelets, Wedgelets, Bandlets, ...lets

Learned dictionaries of patches

[Olshausen and Field, 1997], [Engan et al., 1999], [Lewicki and Sejnowski, 2000], [Aharon et al., 2006], [Roth and Black, 2005], [Lee et al., 2007]

\[
\min_{\alpha_i, D \in \mathbb{C}} \sum_i \frac{1}{2} \| x_i - D \alpha_i \|_2^2 + \lambda \psi(\alpha_i) \\
\]

- \( \psi(\alpha) = \| \alpha \|_0 \) (“\( \ell_0 \) pseudo-norm”)
- \( \psi(\alpha) = \| \alpha \|_1 \) (\( \ell_1 \) norm)
Solving the denoising problem

[Elad and Aharon, 2006]

- Extract all overlapping $8 \times 8$ patches $x_i$.
- Solve a matrix factorization problem:

$$\min_{\alpha_i, D \in \mathcal{C}} \sum_{i=1}^{n} \frac{1}{2} \left\| x_i - D \alpha_i \right\|_2^2 + \lambda \psi(\alpha_i),$$

with $n > 100,000$

- Average the reconstruction of each patch.
Sparse representations for image restoration

K-SVD: [Elad and Aharon, 2006]

Figure: Dictionary trained on a noisy version of the image boat.
Sparse representations for image restoration

Inpainting, Demosaicking

$$\min_{D \in \mathcal{C}, \alpha} \sum_{i} \frac{1}{2} \| \beta_i \otimes (x_i - D \alpha_i) \|^2_2 + \lambda_i \psi(\alpha_i)$$

RAW Image Processing

White balance.
Black substraction.
Conversion to sRGB.
Gamma correction.
Sparse representations for image restoration

[Mairal, Bach, Ponce, Sapiro, and Zisserman, 2009b]
Sparse representations for image restoration

[Mairal, Sapiro, and Elad, 2008d]
Since 1699, when French explorers landed at the great bend of the Mississippi River and celebrated the first Mardi Gras in North America, New Orleans has brewed a fascinating melange of cultures. It was French, then Spanish, then French again, then sold to the United States. Through all these years, and even into the 1900s, others arrived from everywhere: Acadians (Cajuns), Africans, indige-
Sparse representations for image restoration
Inpainting, [Mairal, Elad, and Sapiro, 2008b]
Sparse representations for video restoration

Key ideas for video processing

[Protter and Elad, 2009]

- Using a 3D dictionary.
- Processing of many frames at the same time.
- Dictionary propagation.
Sparse representations for image restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008d]

Figure: Inpainting results.
Sparse representations for image restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008d]

Figure: Inpainting results.
Sparse representations for image restoration
Inpainting, [Mairal, Sapiro, and Elad, 2008d]

Figure: Inpainting results.
Sparse representations for image restoration
Inpainting, [Mairal, Sapiro, and Elad, 2008d]

Figure: Inpainting results.
Sparse representations for image restoration
Inpainting, [Mairal, Sapiro, and Elad, 2008d]

Figure: Inpainting results.
Sparse representations for image restoration
Color video denoising, [Mairal, Sapiro, and Elad, 2008d]

Figure: Denoising results. $\sigma = 25$
Sparse representations for image restoration
Color video denoising, [Mairal, Sapiro, and Elad, 2008d]

Figure: Denoising results. $\sigma = 25$
Sparse representations for image restoration
Color video denoising, [Mairal, Sapiro, and Elad, 2008d]

Figure: Denoising results. $\sigma = 25$
Sparse representations for image restoration
Color video denoising, [Mairal, Sapiro, and Elad, 2008d]

Figure: Denoising results. $\sigma = 25$
Sparse representations for image restoration
Color video denoising, [Mairal, Sapiro, and Elad, 2008d]

Figure: Denoising results. $\sigma = 25$
Digital Zooming
Couzinie-Devy, 2010, Original
Digital Zooming
Couzinie-Devy, 2010, Bicubic
Digital Zooming
Couzinie-Devy, 2010, Proposed method
Digital Zooming
Couzinie-Devy, 2010, Original
Digital Zooming
Couzinie-Devy, 2010, Bicubic
Digital Zooming
Couzinie-Dev, 2010, Proposed approach
Inverse half-toning

Original
Inverse half-toning
Reconstructed image
Inverse half-toning

Original
Inverse half-toning

Reconstructed image
Inverse half-toning

Original
Inverse half-toning

Reconstructed image
Inverse half-toning

Original
Inverse half-toning
Reconstructed image
Inverse half-toning

Original
Inverse half-toning

Reconstructed image
One short slide on compressed sensing

Important message

Sparse coding is not “compressed sensing”.

Compressed sensing is a theory [see Candes, 2006] saying that a sparse signal can be recovered with high probability from a few linear measurements under some conditions.

- Signal Acquisition: \( W^\top x \), where \( W \in \mathbb{R}^{m \times s} \) is a “sensing” matrix with \( s \ll m \).

- Signal Decoding: \( \min_{\alpha \in \mathbb{R}^p} \| \alpha \|_1 \) s.t. \( W^\top x = W^\top D\alpha \).

with extensions to approximately sparse signals, noisy measurements.

Remark

The dictionaries we are using in this lecture do not satisfy the recovery assumptions of compressed sensing.
Important messages

- Patch-based approaches are achieving state-of-the-art results for many image processing task.
- Dictionary Learning adapts to the data you want to restore.
- Dictionary Learning is well adapted to data that admit sparse representation. **Sparsity is for sparse data only.**
Next topics

- Why does the $\ell_1$-norm induce sparsity?
- Some properties of the Lasso.
- Beyond sparsity: Group-sparsity.
- The simplest algorithm for learning dictionaries.
- Links between dictionary learning and matrix factorization techniques.
Why does the $\ell_1$-norm induce sparsity?
Exemple: 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. $\alpha$ 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**:

$$\alpha^* = \text{sign}(x)(|x| - \lambda)^+.$$
Why does the \( \ell_1 \)-norm induce sparsity?

(a) soft-thresholding operator

(b) hard-thresholding operator
Why does the $\ell_1$-norm induce sparsity?

Analysis of the norms in 1D

The gradient of the $\ell_2$-norm vanishes when $\alpha$ get close to 0. On its differentiable part, the norm of the gradient of the $\ell_1$-norm is constant.
Why does the $\ell_1$-norm induce sparsity?

Geometric explanation

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

$$\min_{\alpha \in \mathbb{R}^p} \| x - D\alpha \|^2_2 \quad \text{s.t.} \quad \| \alpha \|_1 \leq T.$$
Important property of the Lasso
Piecewise linearity of the regularization path

Figure: Regularization path of the Lasso
Sparsity-Inducing Norms (1/2)

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

data fitting term

sparsity-inducing norm

Standard approach to enforce sparsity in learning procedures:

- Regularizing by a sparsity-inducing norm \( \psi \).
- The effect of \( \psi \) is to set some \( \alpha_j \)'s to zero, depending on the regularization parameter \( \lambda \geq 0 \).

The most popular choice for \( \psi \):

- The \( \ell_1 \) norm, \( \| \alpha \|_1 = \sum_{j=1}^p |\alpha_j| \).
- For the square loss, Lasso [Tibshirani, 1996].
- However, the \( \ell_1 \) norm encodes poor information, just cardinality!
Another popular choice for $\psi$:

- The $\ell_1$-$\ell_2$ norm,

$$
\sum_{G \in \mathcal{G}} \|\alpha_G\|_2 = \sum_{G \in \mathcal{G}} \left( \sum_{j \in G} \alpha_j^2 \right)^{1/2}, \text{ with } \mathcal{G} \text{ a partition of } \{1, \ldots, p\}.
$$

- The $\ell_1$-$\ell_2$ norm sets to zero groups of non-overlapping variables (as opposed to single variables for the $\ell_1$ norm).
- For the square loss, group Lasso [Yuan and Lin, 2006].
- However, the $\ell_1$-$\ell_2$ norm encodes fixed/static prior information, requires to know in advance how to group the variables!

**Applications:**

- Selecting groups of features instead of individual variables.
- Multi-task learning.
Optimization for Dictionary Learning

\[
\min_{\alpha \in \mathbb{R}^{p \times n}} \sum_{i=1}^{n} \frac{1}{2} \| x_i - D\alpha_i \|^2_2 + \lambda \| \alpha_i \|^1_1
\]

\[ C \triangleq \{ D \in \mathbb{R}^{m \times p} \text{ s.t. } \forall j = 1, \ldots, p, \| d_j \|^2_2 \leq 1 \}. \]

- Classical optimization alternates between \( D \) and \( \alpha \).
- Good results, but slow!
- Instead use online learning [Mairal et al., 2009a]
Optimization for Dictionary Learning
Inpainting a 12-Mpixel photograph

THE SALINAS VALLEY is in Northern California. It is a long narrow valley between two ranges of mountains, and the Salinas River winds and twists up the center until it falls at last into Monterey Bay.

I remember the childhood names for grasses and other flowers. I remember where a toad may live and what time the birds nested. In the summer and what trees and seasons smelled like how people looked and walked and smelled even. The memory of odors is very rich.

I remember that the Sabalina Mountains to the east of the valley were light grey mountains full of sun and loveliness and a kind of invitation. So that you wanted to climb into their warm beeches almost as you want to climb into the lap of a beloved mother. They were beckoning mountains with a brown grass love. The Santa Lucias stood up against the sky to the west and kept the valley from the open sea, and they were dark and brooding-unfriendly and dangerous. I always found in myself a need of west and a love of east. Where I must get such an idea I cannot say, unless it could be that the morning came over the peaks of the Sabalina and the night returned back from the ridges of the Santa Lucias. It may be that the birth and death of the day had some part in my feeling about the two ranges of mountains.

From both sides of the valley little streams slipped out of the high canyons and fell into the bed of the salinas river. In the winter of wet years the streams ran full-freshet, and they swelled the river until sometimes it raged and boiled, bank full, and then it was a destroyer. The river tore the edges of the farm lands and washed whole acres down; it toppled barns and houses into itself, to go floating and bobbing away. It trapped cows and pigs and sheep and drowned them in its muddy brown water and carried them to the sea. Then when the late spring came, the river drew to from its banks and the sand banks appeared. And in the summer the river didn’t run at all above ground. Some pools would be left in the deep swamp places under a high bank. The lilies and grasses grew thick, and willows straightened up with the flood-driven in their upper branches. The Salinas was only a part-time river. The summer sun drove it underground. It was not a first river at all, but it was the only one we had and so we boasted about it how dangerous it was in a wet winter and how dry it was in a dry summer. You can boast about anything if it’s all you have. Maybe the less you have, the more you are required to boast.

The floor of the Salinas Valley, between the ranges and below the foothills, is level because this valley used to be the bottom of a hundred-mile inlet from the sea. The river mouth at Moss Landing was centuries ago the entrance to this long inland water. Once, fifty miles down the valley, my father bored a well. The drill came up first with topsoil and then with gravel and then with white sand sand, rock of shells and even pebbles.
Optimization for Dictionary Learning

Inpainting a 12-Mpixel photograph
Optimization for Dictionary Learning
Inpainting a 12-Mpxel photograph
Optimization for Dictionary Learning
Inpainting a 12-Mpixel photograph
Matrix Factorization Problems and Dictionary Learning

\[
\min_{\alpha \in \mathbb{R}^{p \times n}} \sum_{i=1}^{n} \frac{1}{2} \| x_i - D\alpha_i \|_2^2 + \lambda \| \alpha_i \|_1
\]

can be rewritten

\[
\min_{\alpha \in \mathbb{R}^{p \times n}} \frac{1}{2} \| X - D\alpha \|_F^2 + \lambda \| \alpha \|_1,
\]

where \( X = [x_1, \ldots, x_n] \) and \( \alpha = [\alpha_1, \ldots, \alpha_n] \).
Matrix Factorization Problems and Dictionary Learning

PCA

\[
\min_{\alpha \in \mathbb{R}^{p \times n}} \| X - D\alpha \|_F^2,
\]

with the additional constraints that \( D \) is orthonormal and \( \alpha^T \) is orthogonal.

\( D = [d_1, \ldots, d_p] \) are the principal components.
Matrix Factorization Problems and Dictionary Learning

Hard clustering

\[
\min_{\alpha \in \mathbb{R}^{p \times n}, D \in \mathbb{R}^{m \times p}} \|X - D\alpha\|_F^2,
\]

with the additional constraints that \(\alpha\) is binary and its columns sum to one.
Matrix Factorization Problems and Dictionary Learning

Soft clustering

\[ \min_{\alpha \in \mathbb{R}^{p \times n}, D \in \mathbb{R}^{m \times p}} \| X - D\alpha \|_F^2, \]

with the additional constraints that the columns of \( \alpha \) sum to one.
Matrix Factorization Problems and Dictionary Learning

Non-negative matrix factorization [Lee and Seung, 2001]

\[
\min_{\alpha \in \mathbb{R}^{p \times n}, D \in \mathbb{R}^{m \times p}} \| X - D\alpha \|_F^2,
\]

with the additional constraints that the entries of $D$ and $\alpha$ are non-negative.
Matrix Factorization Problems and Dictionary Learning
NMF+sparsity?

\[
\min_{\alpha \in \mathbb{R}^{p \times n}, D \in \mathbb{R}^{m \times p}} \|X - D\alpha\|_F^2 + \lambda \|\alpha\|_1
\]

with the additional constraints that the entries of \(D\) and \(\alpha\) are non-negative.

**Most of these formulations can be addressed the same types of algorithms.**
Matrix Factorization Problems and Dictionary Learning

Natural Patches

(a) PCA  
(b) NNMF  
(c) DL
Matrix Factorization Problems and Dictionary Learning

Faces

(d) PCA  
(e) NNMF  
(f) DL
Important messages

- The $\ell_1$-norm induces sparsity and shrinks the coefficients (soft-thresholding)
- The regularization path of the Lasso is piecewise linear.
- Sparsity can be induced at the group level.
- Learning the dictionary is simple, fast and scalable.
- Dictionary learning is related to several matrix factorization problems.

**Software SPAMS is available for all of this:**
www.di.ens.fr/willow/SPAMS/.
Next topics: Computer Vision

- Intriguing results on the use of dictionary learning for bags of words.
- Modelling the local appearance of image patches.
Learning Codebooks for Image Classification

Idea

Replacing Vector Quantization by Learned Dictionaries!

- unsupervised: [Yang et al., 2009]
- supervised: [Boureau et al., 2010, Yang et al., 2010]
Learning Codebooks for Image Classification

Let an image be represented by a set of low-level descriptors $x_i$ at $N$ locations identified with their indices $i = 1, \ldots, N$.

- **hard-quantization:**

  $$x_i \approx D\alpha_i, \quad \alpha_i \in \{0, 1\}^p \quad \text{and} \quad \sum_{j=1}^{p} \alpha_i[j] = 1$$

- **soft-quantization:**

  $$\alpha_i[j] = \frac{e^{-\beta\|x_i-d_j\|_2^2}}{\sum_{k=1}^{p} e^{-\beta\|x_i-d_k\|_2^2}}$$

- **sparse coding:**

  $$x_i \approx D\alpha_i, \quad \alpha_i = \arg\min_{\alpha} \frac{1}{2}\|x_i - D\alpha\|_2^2 + \lambda\|\alpha\|_1$$
Yang et al. [2009] have won the PASCAL VOC’09 challenge using this kind of techniques.
Learning dictionaries with a discriminative cost function

Idea:
Let us consider 2 sets $S_-, S_+$ of signals representing 2 different classes. Each set should admit a dictionary best adapted to its reconstruction.

Classification procedure for a signal $x \in \mathbb{R}^n$:

$$\min(R^*(x, D_-), R^*(x, D_+))$$

where

$$R^*(x, D) = \min_{\alpha \in \mathbb{R}^p} \|x - D\alpha\|_2^2 \text{ s.t. } \|\alpha\|_0 \leq L.$$ 

“Reconstructive” training

$$\begin{cases} 
\min_{D_-} \sum_{i \in S_-} R^*(x_i, D_-) \\
\min_{D_+} \sum_{i \in S_+} R^*(x_i, D_+) 
\end{cases}$$

[Grosse et al., 2007], [Huang and Aviyente, 2006], [Sprechmann et al., 2010] for unsupervised clustering (CVPR '10)
“Discriminative” training

[Mairal, Bach, Ponce, Sapiro, and Zisserman, 2008a]

\[
\min_{D_-, D_+} \sum_i C \left( \lambda z_i (R^*(x_i, D_-) - R^*(x_i, D_+)) \right),
\]

where \(z_i \in \{-1, +1\}\) is the label of \(x_i\).
Learning dictionaries with a discriminative cost function
Examples of dictionaries

Top: reconstructive, Bottom: discriminative, Left: Bicycle, Right: Background.
Learning dictionaries with a discriminative cost function
Texture segmentation
Learning dictionaries with a discriminative cost function
Texture segmentation
Learning dictionaries with a discriminative cost function

Pixelwise classification
Learning dictionaries with a discriminative cost function
weakly-supervised pixel classification
Application to edge detection and classification
[Mairal, Leordeanu, Bach, Hebert, and Ponce, 2008c]

Good edges

Bad edges
Application to edge detection and classification
Berkeley segmentation benchmark

Raw edge detection on the right
Application to edge detection and classification
Berkeley segmentation benchmark

Raw edge detection on the right
Application to edge detection and classification
Berkeley segmentation benchmark
Application to edge detection and classification

Contour-based classifier: [Leordeanu, Hebert, and Sukthankar, 2007]

Is there a bike, a motorbike, a car or a person on this image?
Application to edge detection and classification
Application to edge detection and classification
Performance gain due to the prefiltering

<table>
<thead>
<tr>
<th>Category</th>
<th>Ours + [Leordeanu '07]</th>
<th>[Leordeanu '07]</th>
<th>[Winn '05]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>96.8%</td>
<td>89.4%</td>
<td>76.9%</td>
</tr>
</tbody>
</table>

Recognition rates for the same experiment as [Winn et al., 2005] on VOC 2005.

<table>
<thead>
<tr>
<th>Category</th>
<th>Ours + [Leordeanu '07]</th>
<th>[Leordeanu '07]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aeroplane</td>
<td>71.9%</td>
<td>61.9%</td>
</tr>
<tr>
<td>Boat</td>
<td>67.1%</td>
<td>56.4%</td>
</tr>
<tr>
<td>Cat</td>
<td>82.6%</td>
<td>53.4%</td>
</tr>
<tr>
<td>Cow</td>
<td>68.7%</td>
<td>59.2%</td>
</tr>
<tr>
<td>Horse</td>
<td>76.0%</td>
<td>67%</td>
</tr>
<tr>
<td>Motorbike</td>
<td>80.6%</td>
<td>73.6%</td>
</tr>
<tr>
<td>Sheep</td>
<td>72.9%</td>
<td>58.4%</td>
</tr>
<tr>
<td>Tvmonitor</td>
<td>87.7%</td>
<td>83.8%</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>75.9%</strong></td>
<td><strong>64.2%</strong></td>
</tr>
</tbody>
</table>

Recognition performance at equal error rate for 8 classes on a subset of images from Pascal 07.
Digital Art Authentification
Data Courtesy of Hugues, Graham, and Rockmore [2009]

Authentic

Fake
Digital Art Authentification
Data Courtesy of Hugues, Graham, and Rockmore [2009]

Authentic

Fake

Fake
Digital Art Authentification
Data Courtesy of Hugues, Graham, and Rockmore [2009]

Authentic

Fake

Authentic
Important messages

- Learned dictionaries are well adapted to model the local appearance of images and edges.
- They can be used to learn dictionaries of SIFT features.
Next topics

- Optimization for solving sparse decomposition problems
- Optimization for dictionary learning
Recall: The Sparse Decomposition Problem

\[
\min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \| x - D\alpha \|^2_2 + \lambda \psi(\alpha)
\]

\(\psi\) induces sparsity in \(\alpha\). It can be

- the \(\ell_0\) “pseudo-norm”. \(\|\alpha\|_0 \triangleq \# \{ i \; \text{s.t.} \; \alpha[i] \neq 0 \}\) (NP-hard)
- the \(\ell_1\) norm. \(\|\alpha\|_1 \triangleq \sum_{i=1}^p |\alpha[i]|\) (convex)
- \(\ldots\)

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-\(\ell_2\) methods [Daubechies et al., 2009], ...
- active-set methods [Roth and Fischer, 2008].
- ...

Julien Mairal
Sparse Coding and Dictionary Learning

93/137
Matching Pursuit

\[ \alpha = (0, 0, 0) \]
Matching Pursuit

\[ \alpha = (0, 0, 0) \]
Matching Pursuit

\[ \mathbf{d}_2 \]

\[ \mathbf{d}_1 \]

\[ \mathbf{z} \]

\[ \mathbf{r} = \langle \mathbf{r}, \mathbf{d}_3 \rangle \mathbf{d}_3 \]

\[ \mathbf{X} \]

\[ \mathbf{r} = \langle \mathbf{r}, \mathbf{d}_3 \rangle \mathbf{d}_3 \]

\[ \mathbf{d}_3 \]

\[ \mathbf{d}_1 \]

\[ \mathbf{d}_2 \]

\[ \alpha = (0, 0, 0) \]
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

\[ \alpha = (0, 0.24, 0.65) \]
Matching Pursuit

\[
\begin{align*}
\min_{\alpha \in \mathbb{R}^p} & & \| x - D \alpha \|_2^2 \\
\text{s.t.} & & \| \alpha \|_0 \leq L
\end{align*}
\]

1: \( \alpha \leftarrow 0 \)
2: \( r \leftarrow x \) (residual).
3: \textbf{while} \( \| \alpha \|_0 < L \) \textbf{do}
4: \quad \text{Select the atom with maximum correlation with the residual}
5: \qquad \hat{i} \leftarrow \arg \max_{i=1,\ldots,p} |d_i^T r|
6: \quad \text{Update the residual and the coefficients}
7: \qquad \alpha[\hat{i}] \leftarrow \alpha[\hat{i}] + d_{\hat{i}}^T r
8: \qquad r \leftarrow r - (d_{\hat{i}}^T r) d_{\hat{i}}
9: \textbf{end while}
Orthogonal Matching Pursuit

\[ \mathbf{x} = (0, 0, 0) \]

\[ \Gamma = \emptyset \]
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

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

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

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

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

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

6: Update the coefficients

$$\alpha_{\Gamma} \leftarrow (D_{\Gamma}^T D_{\Gamma})^{-1} D_{\Gamma}^T x.$$  

7: end for
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 $G = D^T D$ once in for all,
- Maintain the computation of $D^T r$ for each signal,
- Maintain a Cholesky decomposition of $(D^T \Gamma D \Gamma)^{-1}$ for each signal.

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

$$O(p^2 m) + O(nL^3) + O(n(pm + pL^2)) = O(np(m + L^2))$$

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/

```matlab
>> 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 a 8-cores 2.83Ghz machine: 230000 signals processed per second!
Optimality conditions of the Lasso

Nonsmooth optimization

Directional derivatives and subgradients are useful tools for studying $\ell_1$-decomposition problems:

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

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 \( u \) at \( \alpha \):

\[
\nabla f(\alpha, u) = \lim_{t \to 0^+} \frac{f(\alpha + tu) - f(\alpha)}{t}
\]

- Main idea: in non smooth situations, one may need to look at all directions \( u \) and not simply \( p \) independent ones!

- **Proposition 1**: if \( f \) is differentiable in \( \alpha \), \( \nabla f(\alpha, u) = \nabla f(\alpha)^T u \).

- **Proposition 2**: \( \alpha \) is optimal iff for all \( u \) in \( \mathbb{R}^p \), \( \nabla f(\alpha, u) \geq 0 \).
Optimality conditions of the Lasso

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

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

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

which is equivalent to the following conditions:

$$\forall i = 1, \ldots, p, \quad \left\{ \begin{array}{ll}
|d_i^T (x - D\alpha^*)| & \leq \lambda \\
 d_i^T (x - D\alpha^*) & = \lambda \text{sign}(\alpha^*[i]) \quad \text{if } \alpha^*[i] = 0 \\
\end{array} \right.$$

$$\quad \text{if } \alpha^*[i] \neq 0$$
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 is piecewise linear.
Homotopy, LARS

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

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

The regularization path is piecewise linear:

\[D_{\Gamma}^T(x - D_{\Gamma}\alpha^*_\Gamma) = \lambda \text{sign}(\alpha^*_\Gamma)\]
\[\alpha^*_\Gamma(\lambda) = (D_{\Gamma}^T D_{\Gamma})^{-1}(D_{\Gamma}^T x - \lambda \text{sign}(\alpha^*_\Gamma)) = A + \lambda B\]

A simple interpretation of LARS

- Start from the trivial solution \((\lambda = \|D^T x\|_\infty, \alpha^*(\lambda) = 0)\).
- Maintain the computations of \(|d_i^T(x - D\alpha^*(\lambda))|\) for all \(i\).
- Maintain the computation of the current direction \(B\).
- Follow the path by reducing \(\lambda\) until the next kink.
Example with the software SPAMS

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

```matlab
>> I = double(imread('data/lena.eps'))/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

\[
\min_{\alpha_+, \alpha_-} \frac{1}{2} \|x - D_+\alpha_+ + D_-\alpha_-\|_2^2 + \lambda \alpha_+^T 1 + \lambda \alpha_-^T 1 \quad \text{s.t. } \alpha_-, \alpha_+ \geq 0.
\]

- For this **specific** problem, coordinate descent is **convergent**.
- Supposing \(\|d_i\|_2 = 1\), updating the coordinate \(i\):

\[
\alpha[i] \leftarrow \arg \min_{\beta} \frac{1}{2} \| x - \sum_{j \neq i} \alpha[j]d_j - \beta d_i \|_2^2 + \lambda |\beta| \leftarrow \text{sign}(d_i^T r)(|d_i^T r| - \lambda)^+
\]

\(\Rightarrow\) **soft-thresholding**!
Example with the software SPAMS
http://www.di.ens.fr/willow/SPAMS/

```matlab
>> I=double(imread('data/lena.eps'))/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 continuously differentiable with a Lipschitz gradient.
- Generalize the idea of gradient descent

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

\[
\leftarrow \arg \min_{\alpha \in \mathbb{R}} \frac{1}{2} \| \alpha - (\alpha_k - \frac{1}{L} \nabla f(\alpha_k)) \|_2^2 + \frac{\lambda}{L} \psi(\alpha)
\]

When \(\lambda = 0\), this is equivalent to a classical gradient descent step.
first-order/proximal methods

- They require solving efficiently the proximal operator

\[
\min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \|u - \alpha\|_2^2 + \lambda \psi(\alpha)
\]

- For the $\ell_1$-norm, this amounts to a soft-thresholding:

\[
\alpha^*[i] = \text{sign}(u[i])(u[i] - \lambda)^+. 
\]


- suited for large-scale experiments.
Optimization for Grouped Sparsity

The formulation:

\[
\min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \| x - D\alpha \|_2^2 + \lambda \sum_{g \in G} \| \alpha_g \|_q
\]

- data fitting term
- group-sparsity-inducing regularization

The main class of algorithms for solving grouped-sparsity problems are

- Greedy approaches
- Block-coordinate descent
- Proximal methods
Optimization for Grouped Sparsity

The proximal operator:

\[
\min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \| u - \alpha \|_2^2 + \lambda \sum_{g \in G} \| \alpha_g \|_q
\]

For \( q = 2 \),

\[
\alpha_g^* = \frac{u_g}{\| u_g \|_2} \left( \| u_g \|_2 - \lambda \right)^+, \quad \forall g \in G
\]

For \( q = \infty \),

\[
\alpha_g^* = u_g - \Pi_{\| \cdot \|_1 \leq \lambda} [u_g], \quad \forall g \in G
\]

These formula generalize soft-thresholding to groups of variables. They are used in block-coordinate descent and proximal algorithms.
Reweighted $\ell_2$

Let us start from something simple

$$a^2 - 2ab + b^2 \geq 0.$$  

Then

$$a \leq \frac{1}{2} \left( \frac{a^2}{b} + b \right)$$  

with equality iff $a = b$

and

$$\|\alpha\|_1 = \min_{\eta_j \geq 0} \frac{1}{2} \sum_{j=1}^{p} \frac{\alpha[j]^2}{\eta_j} + \eta_j.$$  

The formulation becomes

$$\min_{\alpha, \eta_j \geq \varepsilon} \frac{1}{2} \|x - D\alpha\|_2^2 + \frac{\lambda}{2} \sum_{j=1}^{p} \frac{\alpha[j]^2}{\eta_j} + \eta_j.$$
Important messages

- Greedy methods directly address the NP-hard $\ell_0$-decomposition problem.
- 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 $\ell_2$ methods, stochastic variants, variants of OMP, . . .
Optimization for Dictionary Learning

\[
\min_{\alpha \in \mathbb{R}^{p \times n}, D \in \mathcal{C}} \sum_{i=1}^{n} \frac{1}{2} \|x_i - D\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1
\]

\[\mathcal{C} \triangleq \{D \in \mathbb{R}^{m \times p} \text{ s.t. } \forall j = 1, \ldots, p, \|d_j\|_2 \leq 1\} \]

- Classical optimization alternates between \(D\) and \(\alpha\).
- Good results, but very slow!
Optimization for Dictionary Learning

[Mairal, Bach, Ponce, and Sapiro, 2009a]

Classical formulation of dictionary learning

\[
\min_{D \in C} f_n(D) = \min_{D \in C} \frac{1}{n} \sum_{i=1}^{n} l(x_i, D),
\]

where

\[
l(x, D) \triangleq \min_{\alpha} \frac{1}{2} \|x - D\alpha\|_2^2 + \lambda \|\alpha\|_1.
\]

Which formulation are we interested in?

\[
\min_{D \in C} \left\{ f(D) = \mathbb{E}_x[l(x, D)] \approx \lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} l(x_i, D) \right\}
\]

[Bottou and Bousquet, 2008]: Online learning can

- handle potentially infinite or dynamic datasets,
- be dramatically faster than batch algorithms.
Optimization for Dictionary Learning

Require: \( D_0 \in \mathbb{R}^{m \times p} \) (initial dictionary); \( \lambda \in \mathbb{R} \)

1: \( A_0 = 0, \ B_0 = 0 \).
2: for \( t=1,\ldots,T \) do
3: Draw \( x_t \)
4: Sparse Coding

\[ \alpha_t \leftarrow \arg \min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \| x_t - D_{t-1} \alpha \|_2^2 + \lambda \| \alpha \|_1, \]

5: Aggregate sufficient statistics
\( A_t \leftarrow A_{t-1} + \alpha_t \alpha_t^T, \ B_t \leftarrow B_{t-1} + x_t \alpha_t^T \)
6: Dictionary Update (block-coordinate descent)

\[ D_t \leftarrow \arg \min_{D \in \mathcal{C}} \frac{1}{t} \sum_{i=1}^{t} \left( \frac{1}{2} \| x_i - D \alpha_i \|_2^2 + \lambda \| \alpha_i \|_1 \right). \]

7: end for
Optimization for Dictionary Learning

Which guarantees do we have?
Under a few reasonable assumptions,
- we build a surrogate function \( \hat{f}_t \) of the expected cost \( f \) verifying

\[
\lim_{t \to +\infty} \hat{f}_t(D_t) - f(D_t) = 0,
\]

- \( D_t \) is asymptotically close to a stationary point.

Extensions (all implemented in SPAMS)
- non-negative matrix decompositions.
- sparse PCA (sparse dictionaries).
- fused-lasso regularizations (piecewise constant dictionaries)
Optimization for Dictionary Learning
Experimental results, batch vs online

\[ m = 8 \times 8, \ p = 256 \]
Optimization for Dictionary Learning
Experimental results, batch vs online

\[ m = 12 \times 12 \times 3, \; p = 512 \]
References I


References II


References VI


References VII


