# Backpropagation Rules for Sparse Coding (Task-Driven Dictionary Learning)

Julien Mairal

UC Berkeley

Edinburgh, ICML, June 2012

# Other Persons Involved

Francis Bach

Jean Ponce

Florent Couzinie-Devy

INRIA - Willow and Sierra Teams

## References


[1] J. Mairal, F. Bach and J. Ponce. Task-Driven Dictionary Learning. PAMI. 2012;

[2] F. Couzinie-Devy, J. Mairal, F. Bach and J. Ponce. Dictionary Learning for Deblurring and Digital Zoom. arXiv:1110.0957. 2011.

## What this work is about

- a few attempts of **supervised feature learning**;
- dictionary learning adapted to other tasks than reconstruction;
- (some links between sparse coding and neural networks).

## Applications

- nonlinear inverse image problems;
- digits/patch/image classification;

# Outline

1. Quick Introduction to Dictionary Learning

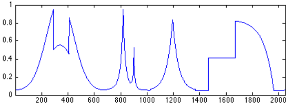
2. Two Layers Model for Regression and Classification

3. Applications

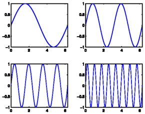
# Outline

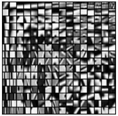
# What is a Sparse Linear Model?

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




Let $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_p] \in \mathbb{R}^{m \times p}$ 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** $\boldsymbol{\alpha}$ in $\mathbb{R}^p$ such that $\mathbf{x} \approx \mathbf{D}\boldsymbol{\alpha}$. We call $\boldsymbol{\alpha}$ the **sparse code**.

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

# The Sparse Decomposition Problem

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

$\psi$ induces sparsity in $\boldsymbol{\alpha}$. It can be

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

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]

# The Dictionary Learning Problem

Given training signals $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, e.g., natural image patches

$$\min_{\boldsymbol{\alpha}_i, \mathbf{D} \in \mathcal{D}} \sum_i \underbrace{\frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2}_{\text{reconstruction}} + \underbrace{\lambda\psi(\boldsymbol{\alpha}_i)}_{\text{sparsity}}$$

Originally introduced by Olshausen and Field [1996].

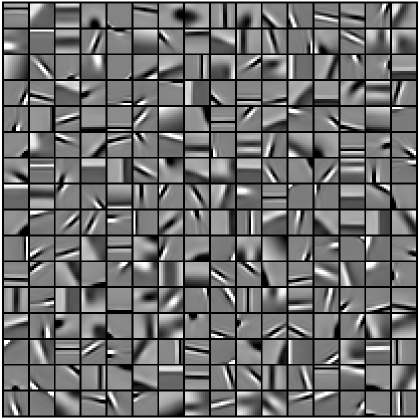
The matrix factorization view

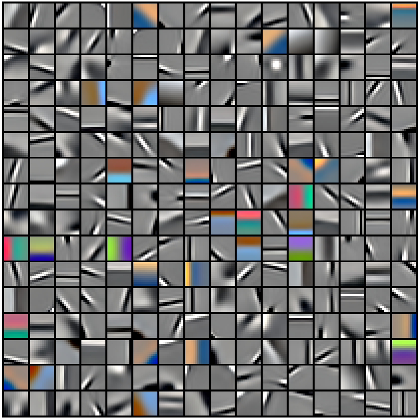
$$\min_{\mathbf{A} \in \mathbb{R}^{p \times n}, \mathbf{D} \in \mathcal{D}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_\mathsf{F}^2 + \lambda\psi(\mathbf{A}).$$

Other related matrix factorization problems: vector quantization, non-negative matrix factorization, principal component analysis, probabilistic topic models, independent component analysis...

# Dictionary Learning of Natural Image Patches

## Grayscale vs color image patches





# Sparse representations for image restoration

## Solving the denoising problem

[Elad and Aharon, 2006]

- Extract all overlapping $8 \times 8$ patches $\mathbf{y}_i$.
- Solve a matrix factorization problem:

$$\min_{\boldsymbol{\alpha}_i, \mathbf{D} \in \mathcal{D}} \sum_{i=1}^{n} \underbrace{\frac{1}{2}\|\mathbf{y}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2}_{\text{reconstruction}} + \underbrace{\lambda\psi(\boldsymbol{\alpha}_i)}_{\text{sparsity}},$$

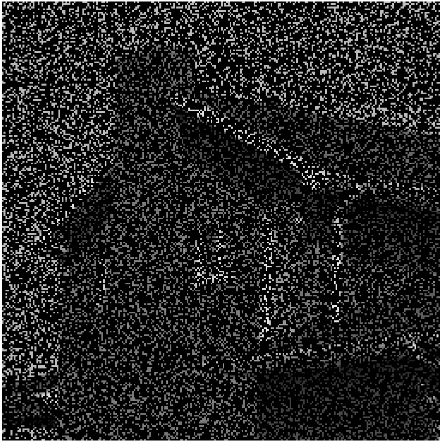
with $n > 100,000$

- Average the reconstruction of each patch.

# Dictionary Learning for Image Restoration

Denoising: [Elad and Aharon, 2006]

Inpainting: [Mairal, Sapiro, and Elad, 2008c]



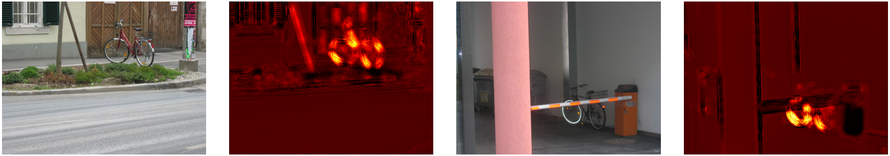
# Dictionary Learning for Image Restoration

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

# Previous work: Learning Discriminative Dictionaries
[Mairal et al., 2008b]

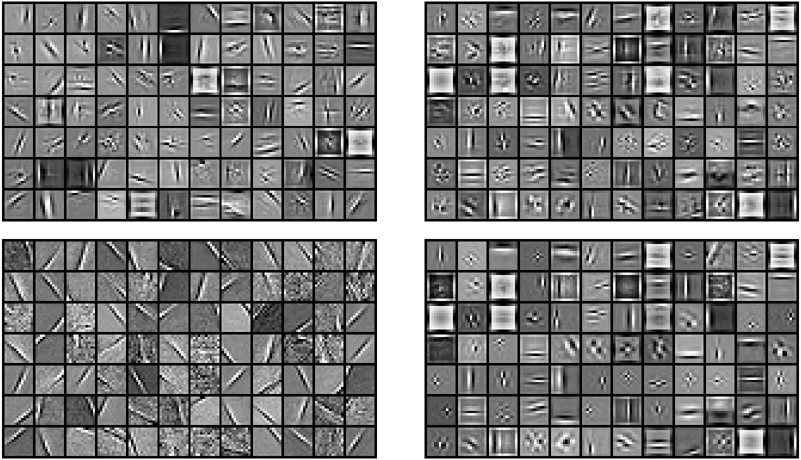
## Dictionaries can be tuned for classification tasks

- learns the local appearance of objects, textures and edges in images.
- heuristic optimization.



# Previous work: Learning Discriminative Dictionaries

## Examples of dictionaries

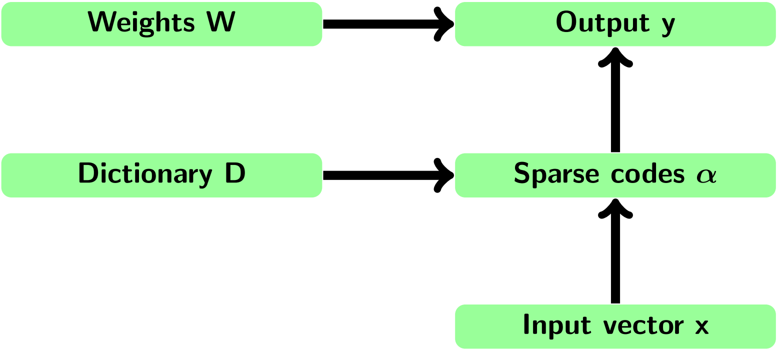


Top: reconstructive; bottom: discriminative; left: bicycle; right: background.

# Outline

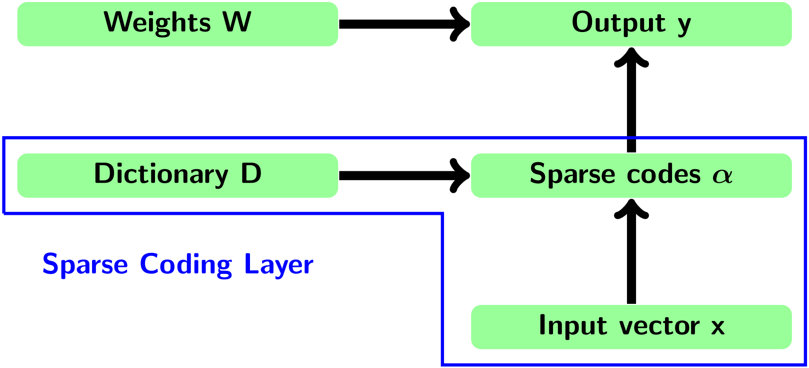
# Two Layers Models

**Use the sparse codes $\alpha$ as feature representation.**
[Raina et al., 2007, Mairal et al., 2008b, Bradley and Bagnell, 2008]



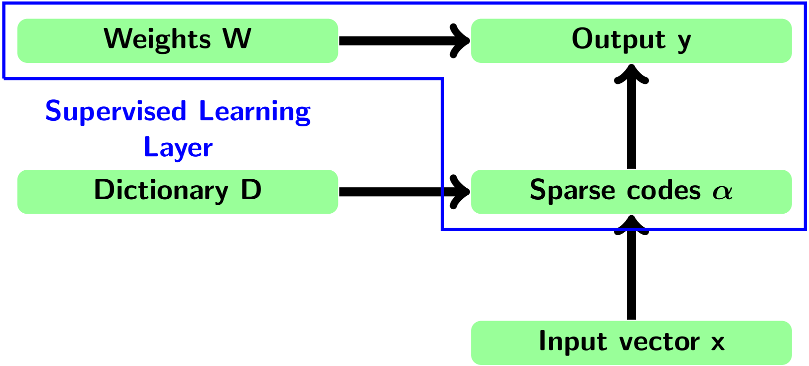
# Two Layers Models

**Use the sparse codes $\alpha$ as feature representation.**
[Raina et al., 2007, Mairal et al., 2008b, Bradley and Bagnell, 2008]



# Two Layers Models

**Use the sparse codes $\alpha$ as feature representation.**
[Raina et al., 2007, Mairal et al., 2008b, Bradley and Bagnell, 2008]



# Two Layers Models

Given a training set $(\mathbf{x}_i, \mathbf{y}_i)_{i=1,\ldots,n}$,

First layer: Dictionary Learning

$$\min_{\boldsymbol{\alpha}_i, \mathbf{D} \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1.$$

This is an unsupervised learning formulation.

Second layer: Supervised Learning

$$\min_{\mathbf{W}} \frac{1}{n} \sum_{i=1}^{n} L(\mathbf{y}_i, \mathbf{W}\boldsymbol{\alpha}_i) + \frac{\gamma}{2} \|\mathbf{W}\|_{\mathrm{F}}^2,$$

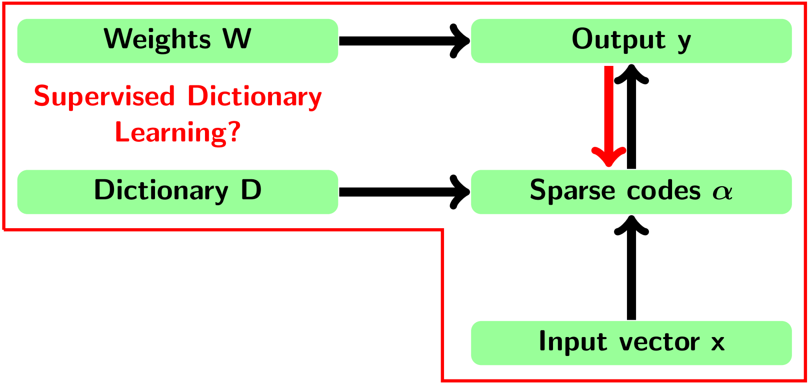
$L$ is an appropriate loss function (often convex).

# Two Layers Models

- **Unsupervised Feature Learning is often suboptimal for supervised learning tasks**

# Two Layers Models

- **Unsupervised Feature Learning is often suboptimal for supervised learning tasks**
- ...but supervised feature learning is hard.



# Supervised Feature Learning in the Literature

- **backpropagation** in the neural network literature (70's); [see LeCun et al., 1998];
- supervised **fine-tuning** of convolutional neural networks, deep networks, restricted boltzmann machines;
- supervised topic models [Blei and McAuliffe, 2008];
- supervision in sparse coding formulations [Mairal et al., 2008a,b, 2012, Bradley and Bagnell, 2008, Boureau et al., 2010, Yang et al., 2010b],...;
- ...

**How do we build a backpropagation rule for dictionary learning?**

Note that Bradley and Bagnell [2008] already use the terminology "backpropagation" for sparse coding.

# Backpropagation rule for sparse coding

Original Formulation:

$$\min_{\mathbf{W},\mathbf{D}\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} L(\mathbf{y}_i, \mathbf{W}\boldsymbol{\alpha}^\star(\mathbf{x}_i, \mathbf{D})) + \frac{\gamma}{2}\|\mathbf{W}\|_\mathsf{F}^2,$$

where

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

Other techniques used for classification tasks

- Bradley and Bagnell [2008]: smooth approximation + implicit differentiation + gradient descent.
- Boureau et al. [2010], Yang et al. [2010b]: heuristic implicit differentiation + gradient descent.

# Backpropagation rule for sparse coding

Formulation with expected cost

$$\min_{\mathbf{W},\mathbf{D}\in\mathcal{D}} \underbrace{\mathbb{E}_{\mathbf{y},\mathbf{x}}[L(\mathbf{y},\mathbf{W}\boldsymbol{\alpha}^\star(\mathbf{x},\mathbf{D}))]}_{f(\mathbf{D},\mathbf{W})} + \frac{\gamma}{2}\|\mathbf{W}\|_F^2,$$

where

$$\boldsymbol{\alpha}^\star(\mathbf{x},\mathbf{D}) \triangleq \operatorname*{arg\,min}_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{\alpha}\|_2^2.$$

the elastic-net [Zou and Hastie, 2005] brings more stability than $\ell_1$ and ensures some regularity (Lipschitz continuity) of the function $\boldsymbol{\alpha}^\star$.

# Main Result

## Differentiability and gradients of $f$

Under a few technical assumptions (the probability distribution of $(\mathbf{y}, \mathbf{x})$ admits a continuous density with compact support, $L$ is twice differentiable), the function $f$ is **differentiable**, and

$$\begin{cases} \nabla_{\mathbf{W}} f(\mathbf{D}, \mathbf{W}) = \mathbb{E}_{\mathbf{y},\mathbf{x}}[\nabla L(\mathbf{y}, \mathbf{W}\boldsymbol{\alpha}^\star)\boldsymbol{\alpha}^{\star\top}], \\ \nabla_{\mathbf{D}} f(\mathbf{D}, \mathbf{W}) = \mathbb{E}_{\mathbf{y},\mathbf{x}}[-\mathbf{D}\boldsymbol{\beta}^\star\boldsymbol{\alpha}^{\star\top} + (\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}^\star)\boldsymbol{\beta}^{\star\top}], \end{cases}$$

and $\boldsymbol{\beta}^\star$ is a vector in $\mathbb{R}^p$ that depends on $\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{D}$ with

$$\boldsymbol{\beta}^\star_{\Lambda^C} = 0 \quad \text{and} \quad \boldsymbol{\beta}^\star_\Lambda = (\mathbf{D}_\Lambda^\top \mathbf{D}_\Lambda + \lambda_2 \mathbf{I})^{-1} \mathbf{W}_\Lambda^\top \nabla L(\mathbf{y}, \mathbf{W}\boldsymbol{\alpha}^\star),$$

where $\Lambda$ denotes the indices of the nonzero coefficients of $\boldsymbol{\alpha}^\star$.

# Main Result

## Differentiability and gradients of $f$

Under a few technical assumptions (the probability distribution of $(\mathbf{y}, \mathbf{x})$ admits a continuous density with compact support, $L$ is twice differentiable), the function $f$ is **differentiable**, and

$$\begin{cases} \nabla_{\mathbf{W}} f(\mathbf{D}, \mathbf{W}) = \mathbb{E}_{\mathbf{y},\mathbf{x}}[\nabla L(\mathbf{y}, \mathbf{W}\boldsymbol{\alpha}^\star)\boldsymbol{\alpha}^{\star\top}], \\ \nabla_{\mathbf{D}} f(\mathbf{D}, \mathbf{W}) = \mathbb{E}_{\mathbf{y},\mathbf{x}}[-\mathbf{D}\boldsymbol{\beta}^\star\boldsymbol{\alpha}^{\star\top} + (\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}^\star)\boldsymbol{\beta}^{\star\top}], \end{cases}$$

and $\boldsymbol{\beta}^\star$ is a vector in $\mathbb{R}^p$ that depends on $\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{D}$ with

$$\boldsymbol{\beta}^\star_{\Lambda^C} = 0 \quad \text{and} \quad \boldsymbol{\beta}^\star_{\Lambda} = (\mathbf{D}_\Lambda^\top \mathbf{D}_\Lambda + \lambda_2 \mathbf{I})^{-1} \mathbf{W}_\Lambda^\top \nabla L(\mathbf{y}, \mathbf{W}\boldsymbol{\alpha}^\star),$$

where $\Lambda$ denotes the indices of the nonzero coefficients of $\boldsymbol{\alpha}^\star$.

$\Longrightarrow$ **stochastic gradient descent**

# Practical Implementation

Learning rule:

$$\mathbf{W} \leftarrow \mathbf{W} - \rho_t\big(\nabla L(\mathbf{y}_t, \mathbf{W}\boldsymbol{\alpha}_t^\star)\boldsymbol{\alpha}_t^{\star\top} + \gamma\mathbf{W}\big),$$
$$\mathbf{D} \leftarrow \Pi_{\mathcal{D}}\Big[\mathbf{D} - \rho_t\big( - \mathbf{D}\boldsymbol{\beta}_t^\star\boldsymbol{\alpha}_t^{\star\top} + (\mathbf{x}_t - \mathbf{D}\boldsymbol{\alpha}_t^\star)\boldsymbol{\beta}_t^{\star\top}\big)\Big],$$

A few tricks

- use mini-batches;
- initialize with unsupervised dictionary learning [Mairal et al., 2010];
- try different learning steps for a few iterations before choosing one;
- rescale the data;
- use homotopy algorithm to compute $\boldsymbol{\alpha}^\star$, and get $\boldsymbol{\beta}^\star$ for free;
- first try $\lambda_2 = 0$, if the algorithm diverges, use $\lambda_2 > 0$.

**see the backpropagation literature [LeCun et al., 1998]**

# Outline

# Application - Multivariate Regression

## Problem

Signals $\mathbf{x}_i$ from an input space $\mathcal{X}$ are associated to transformed signals $\mathbf{y}_i$ from an output space $\mathcal{Y}$ and we want to learn the inverse transformation.

## Formulation

$$\min_{\mathbf{W},\mathbf{D}\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\|\mathbf{y}_i - \mathbf{W}\alpha^{\star}(\mathbf{x}_i,\mathbf{D})\|_2^2$$

## Interpretation

**D** and **W** can be interpreted as linked dictionaries, one in the input space of the $\mathbf{x}_i$'s, one in the output space of the $\mathbf{y}_i$'s.

**Image reconstruction with a patch-based approach.**

# Inverse half-toning

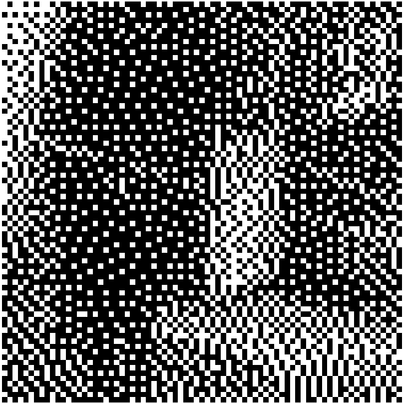
## Original

# Inverse half-toning

## Reconstructed image

# Inverse half-toning

## Reconstructed image



# Linked Dictionaries

## Without Backpropagation

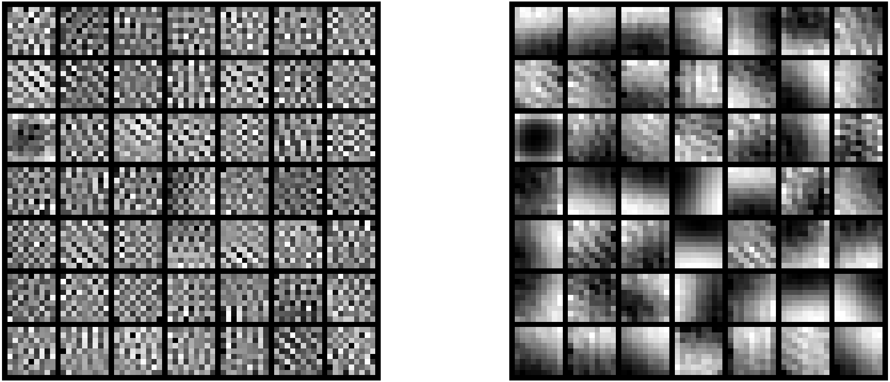
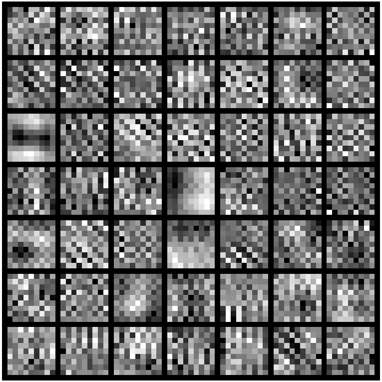


Figure: Left: **D**; right: **W**

# Linked Dictionaries

## With Backpropagation



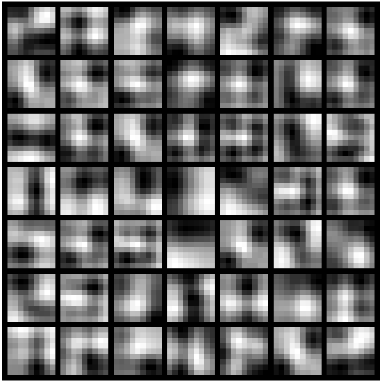
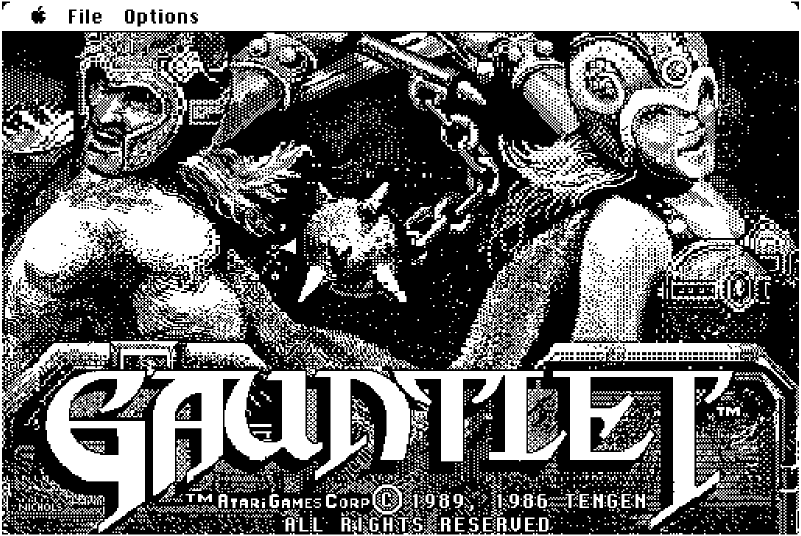


Figure: Left: **D**; right: **W**

# 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

Table: Inverse halftoning. Results are in PSNR. SA-DCT refers to [Dabov et al., 2006], LPA-ICI to [Foi et al., 2004], FIHT2 to [Kite et al., 2000] and WInHD to [Neelamani et al., 2009].

| | Test set | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Image | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| FIHT2 | 24.5 | 28.6 | 29.5 | 28.2 | 29.3 | 26.0 | 25.2 | 24.7 |
| WInHD | 25.7 | 29.2 | 29.4 | 28.7 | 29.4 | 28.1 | 25.6 | 26.4 |
| LPA-ICI | 25.6 | 29.7 | 30.0 | 29.2 | 30.1 | 28.3 | 26.0 | 27.2 |
| SA-DCT | **27.0** | 30.1 | 30.2 | 29.8 | 30.3 | 28.5 | **26.2** | 27.6 |
| Ours | 26.6 | **30.2** | **30.5** | **29.9** | **30.4** | **29.0** | **26.2** | **28.0** |

# Deblurring and Superresolution

Couzinie-Devy et al., 2011

## Brief Summary

- jointly learns low- and high-res dictionaries [Yang et al., 2010a, Zeyde et al., 2012] but brings backpropagation to these approaches;
- combines linear and non-linear models;
- competitive with the state of the art for non-blind image deblurring and image superresolution;

# Digital Zooming

[Couzinie-Devy et al., 2011], Original

# Digital Zooming

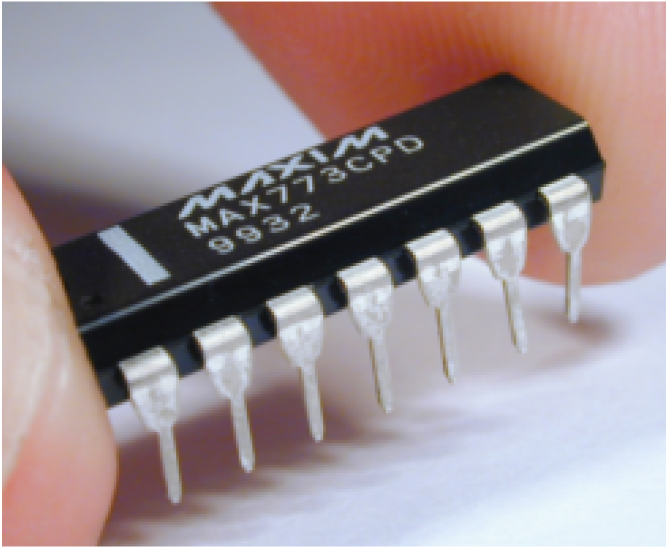
[Couzinie-Devy et al., 2011], Bicubic

# Digital Zooming

[Couzinie-Devy et al., 2011], Result

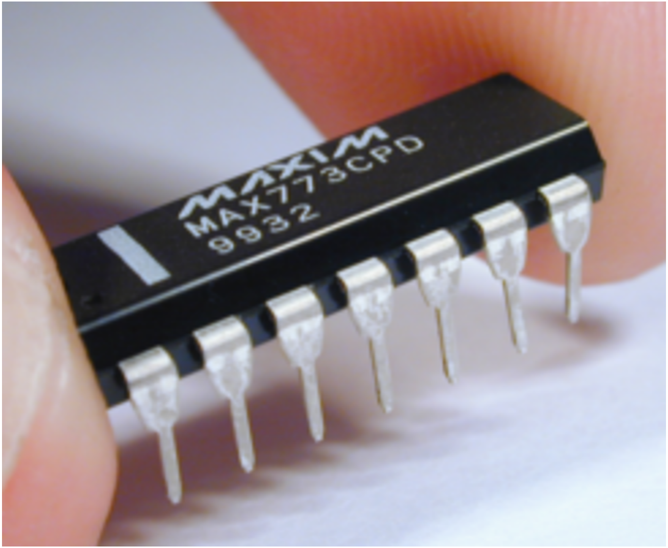
# Digital Zooming

[Couzinie-Devy et al., 2011], Original



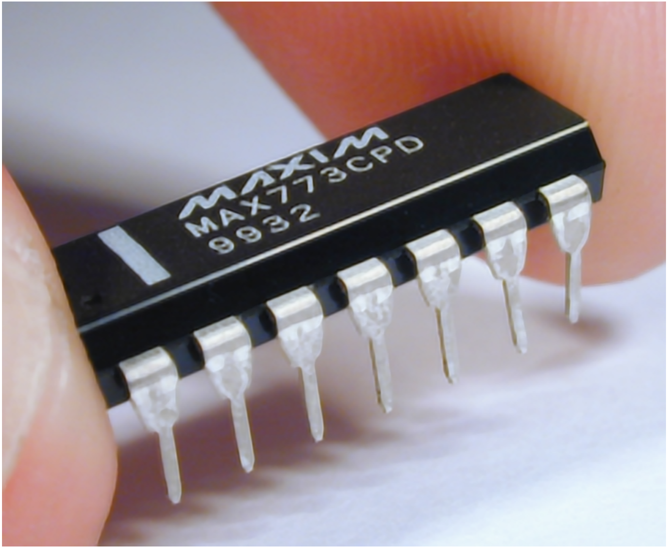
# Digital Zooming

[Couzinie-Devy et al., 2011], Bicubic



# Digital Zooming

[Couzinie-Devy et al., 2011], Result



# Image Deblurring

[Couzinie-Devy et al., 2011], Original

# Image Deblurring

[Couzinie-Devy et al., 2011], Blurry and Noisy

# Image Deblurring

[Couzinie-Devy et al., 2011], Result

# Image Deblurring

[Couzinie-Devy et al., 2011], Original

# Image Deblurring

[Couzinie-Devy et al., 2011], Blurry and Noisy

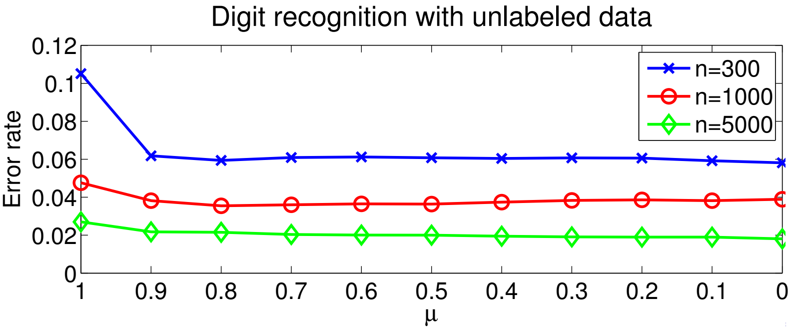
# Image Deblurring

[Couzinie-Devy et al., 2011], Result

# Application - Classification - Digit Recognition

| **D** | unsupervised | | | | supervised | | | |
|---|---|---|---|---|---|---|---|---|
| $p$ | 50 | 100 | 200 | 300 | 50 | 100 | 200 | 300 |
| MNIST | 5.27 | 3.92 | 2.95 | 2.36 | .96 | .73 | .57 | .54 |
| USPS | 8.02 | 6.03 | 5.13 | 4.58 | 3.64 | 3.09 | 2.88 | 2.84 |

Table: Results similar to Ranzato et al. [2007] for MNIST. We extend the training set with shifted versions of the digits by one pixel.



# A few Empirical Conclusions

## Advantages

- in some cases, backpropagation (fine-tuning) significantly improves the prediction performance;
- achieves better or same performance with smaller dictionary sizes (implies faster prediction at test time);

## Drawbacks

- non-convex;
- learning is more difficult;
- in some cases, the prediction performance does not improve.

**This approach would benefit from good novel heuristics for automatically choosing the learning rate.**

# Advertisement SPAMS toolbox (open-source)

- C++ interfaced with **Matlab, R, Python**.
- proximal gradient methods for $\ell_0$, $\ell_1$, **elastic-net, fused-Lasso, group-Lasso, tree group-Lasso, tree-**$\ell_0$**, sparse group Lasso, overlapping group Lasso, trace norm...**
- ...for **square, logistic, multi-class logistic** loss functions.
- handles sparse matrices, provides duality gaps.
- fast implementations of **OMP** and **LARS - homotopy**.
- dictionary learning and matrix factorization (NMF, sparse PCA).
- coordinate descent, block coordinate descent algorithms.
- fast projections onto some convex sets.

**Try it!** `http://www.di.ens.fr/willow/SPAMS/`

# References I


D. Blei and J. McAuliffe. Supervised topic models. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, pages 121–128. MIT Press, 2008.

Y-L. Boureau, F. Bach, Y. Lecun, and J. Ponce. Learning mid-level features for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

D. M. Bradley and J. A. Bagnell. Differentiable sparse coding. In *Advances in Neural Information Processing Systems*. 2008.

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

F. Couzinie-Devy, J. Mairal, F. Bach, and J. Ponce. Dictionary learning for deblurring and digital zoom. *preprint arXiv:1110.0957*, 2011.

# References II

K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Inverse halftoning by pointwise shape-adaptive DCT regularized deconvolution. In *Proceedings of the International TICSP Workshop on Spectral Methods Multirate Signal Processing (SMMSP)*, 2006.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 54(12):3736–3745, December 2006.

A. Foi, V. Katkovnik, K. Egiazarian, and J. Astola. Inverse halftoning based on the anisotropic lpa-ici deconvolution. In *Proceedings of Int. TICSP Workshop Spectral Meth. Multirate Signal Process.*, 2004.

T. D. Kite, N. Damera-Venkata, B. L. Evans, and A. C. Bovik. A fast, high-quality inverse halftoning algorithm for error diffused halftones. *IEEE Transactions on Image Processing*, 9(9):1583–1592, 2000.

# References III

Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop. In G. Orr and Muller K., editors, *Neural Networks: Tricks of the trade*. Springer, 1998.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008a.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *Advances in Neural Information Processing Systems*. 2008b.

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

# References IV

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

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 2010.

J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

R. Neelamani, R.D. Nowak, and R.G. Baraniuk. WInHD: Wavelet-based inverse halftoning via deconvolution. *Rejecta Mathematica*, 1(1): 84–103, 2009.

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.

# References V

R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2007.

M. Ranzato, F. Huang, Y. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

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

J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 19 (11):2861–2873, 2010a.

J. Yang, K. Yu, , and T. Huang. Supervised translation-invariant sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010b.

# References VI

R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. *Curves and Surfaces*, pages 711–730, 2012.

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.