Online Dictionary Learning for Sparse Coding

Julien Mairal[1] Francis Bach[1]
Jean Ponce[2] Guillermo Sapiro[3]

[1]INRIA–Willow project [2]Ecole Normale Supérieure
[3]University of Minnesota

ICML, Montréal, June 2008

What this talk is about

- Learning efficiently dictionaries (basis set) for sparse coding.
- Solving a large-scale matrix factorization problem.
- Making some large-scale image processing problems tractable.
- Proposing an algorithm which extends to NMF, sparse PCA,...

1. The Dictionary Learning Problem
2. Online Dictionary Learning
3. Extensions

1. The Dictionary Learning Problem

2. Online Dictionary Learning

3. Extensions

The Dictionary Learning Problem

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

The Dictionary Learning Problem
[Elad & Aharon ('06)]

Solving the denoising problem

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

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

with $n > 100,000$

- Average the reconstruction of each patch.

The Dictionary Learning Problem
[Mairal, Bach, Ponce, Sapiro & Zisserman ('09)]

Denoising result

The Dictionary Learning Problem
[Mairal, Sapiro & Elad ('08)]

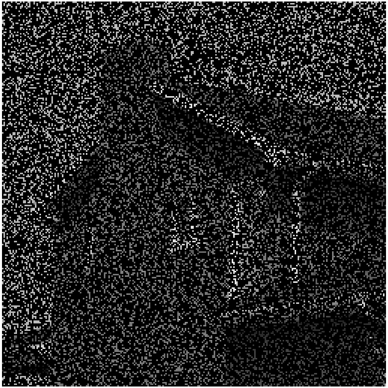
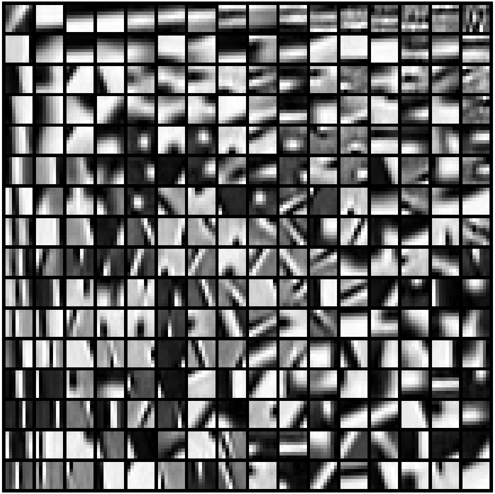


Image completion example

The Dictionary Learning Problem

What does **D** look like?



The Dictionary Learning Problem

$$\min_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^{k \times n} \\ \mathbf{D} \in \mathcal{C}}} \sum_{i=1}^{n} \frac{1}{2} ||\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i||_2^2 + \lambda ||\boldsymbol{\alpha}_i||_1$$

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

- Classical optimization alternates between $\mathbf{D}$ and $\boldsymbol{\alpha}$.
- Good results, but very slow!

1 The Dictionary Learning Problem

2 Online Dictionary Learning

3 Extensions

Online Dictionary Learning

Classical formulation of dictionary learning

$$\min_{\mathbf{D}\in\mathcal{C}} f_n(\mathbf{D}) = \min_{\mathbf{D}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} l(\mathbf{x}_i, \mathbf{D}),$$

where

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

Online Dictionary Learning

Which formulation are we interested in?

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

Online Dictionary Learning

Online learning can

- handle potentially infinite datasets,
- adapt to dynamic training sets,
- be dramatically faster than batch algorithms [Bottou & Bousquet ('08)].

Online Dictionary Learning

Proposed approach

for t=1,...,T **do**

  Draw $\mathbf{x}_t$

  Sparse Coding

$$\boldsymbol{\alpha}_t \leftarrow \underset{\boldsymbol{\alpha}\in\mathbb{R}^k}{\arg\min} \frac{1}{2}||\mathbf{x}_t - \mathbf{D}_{t-1}\boldsymbol{\alpha}||_2^2 + \lambda||\boldsymbol{\alpha}||_1,$$

  Dictionary Learning

$$\mathbf{D}_t \leftarrow \underset{\mathbf{D}\in\mathcal{C}}{\arg\min} \frac{1}{t}\sum_{i=1}^{t}\left(\frac{1}{2}||\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i||_2^2 + \lambda||\boldsymbol{\alpha}_i||_1\right),$$

end for

Implementation details

- Use LARS for the sparse coding step,
- Use a block-coordinate approach for the dictionary update, with warm restart,
- Use a mini-batch.

Which guarantees do we have?

Under a few reasonable assumptions,

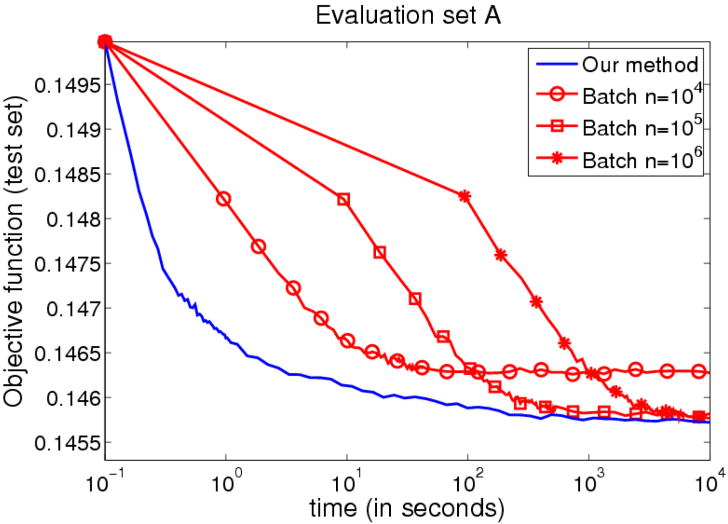
- we build a surrogate function $\hat{f}_t$ of the expected cost f verifying

$$\lim_{t \to +\infty} \hat{f}_t(\mathbf{D}_t) - f(\mathbf{D}_t) = 0,$$

- $\mathbf{D}_t$ is asymptotically close to a stationary point.

Online Dictionary Learning

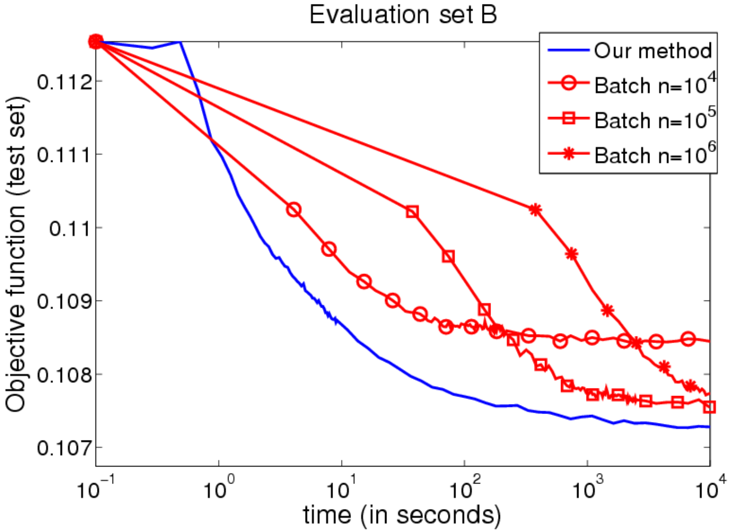
Experimental results, batch vs online



$m = 8 \times 8,\ k = 256$

Online Dictionary Learning

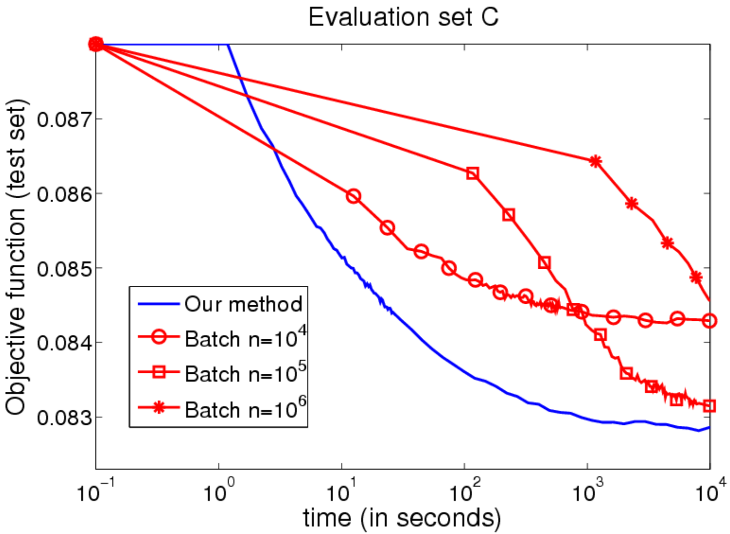
Experimental results, batch vs online



$$m = 12 \times 12 \times 3, \; k = 512$$

Online Dictionary Learning

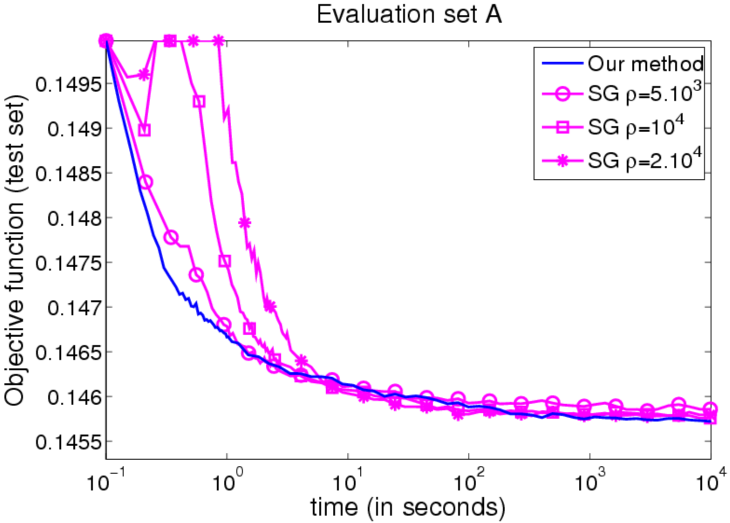
Experimental results, batch vs online



$$m = 16 \times 16,\ k = 1024$$

Online Dictionary Learning

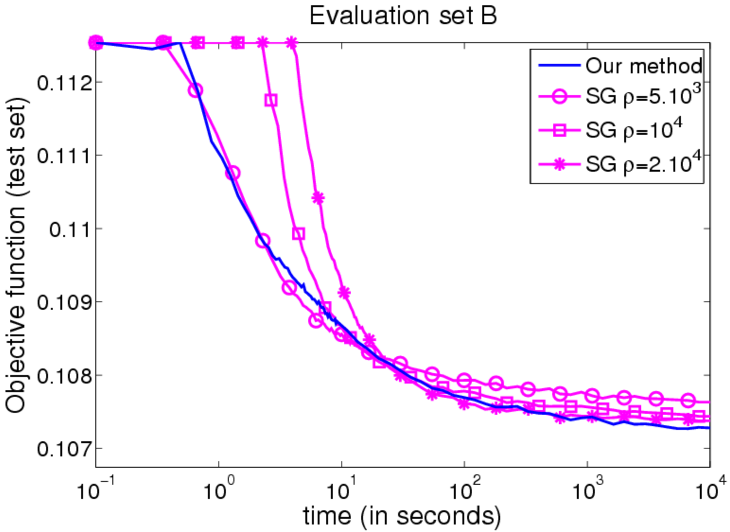
Experimental results, ODL vs SGD



$m = 8 \times 8, \ k = 256$

Online Dictionary Learning

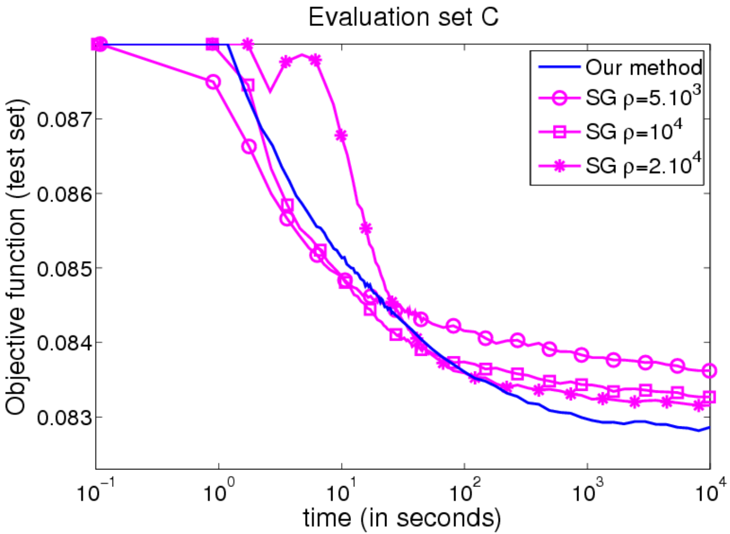
Experimental results, ODL vs SGD



$$m = 12 \times 12 \times 3, \; k = 512$$

Online Dictionary Learning

Experimental results, ODL vs SGD



$$m = 16 \times 16,\ k = 1024$$

Online Dictionary Learning

Inpainting a 12-Mpixel photograph

Online Dictionary Learning

Inpainting a 12-Mpixel photograph

g mountain
alley from th
lf a dread o
orning cam

Online Dictionary Learning
Inpainting a 12-Mpixel photograph

1. The Dictionary Learning Problem
2. Online Dictionary Learning
3. Extensions

Extension to NMF and sparse PCA

NMF extension

$$\min_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^{k \times n} \\ \mathbf{D} \in \mathcal{C}}} \sum_{i=1}^{n} \frac{1}{2} ||\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i||_2^2 \ \text{ s.t. } \ \boldsymbol{\alpha}_i \geq 0, \quad \mathbf{D} \geq 0.$$

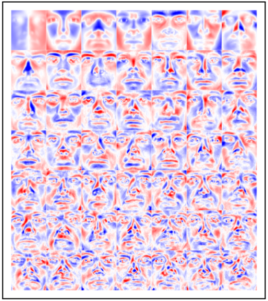
SPCA extension

$$\min_{\substack{\boldsymbol{\alpha} \in \mathbb{R}^{k \times n} \\ \mathbf{D} \in \mathcal{C}'}} \sum_{i=1}^{n} \frac{1}{2} ||\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i||_2^2 + \lambda ||\boldsymbol{\alpha}_1||_1$$

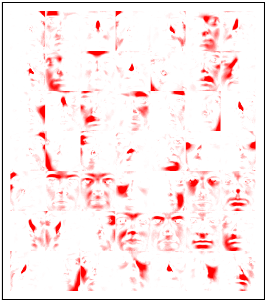
$$\mathcal{C}' \triangleq \{\mathbf{D} \in \mathbb{R}^{m \times k} \ \text{ s.t. } \ \forall j \ \ ||\mathbf{d}_j||_2^2 + \gamma ||\mathbf{d}_j||_1 \leq 1\}.$$

Extension to NMF and sparse PCA

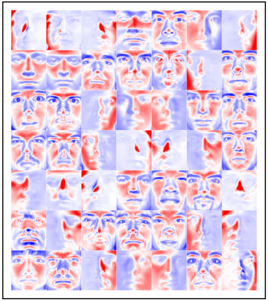
Faces: Extended Yale Database B



(a) PCA



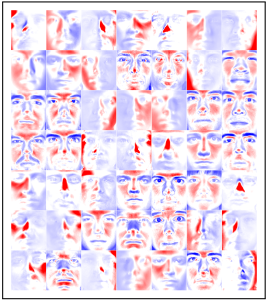
(b) NNMF



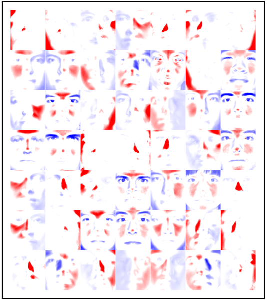
(c) DL

Extension to NMF and sparse PCA

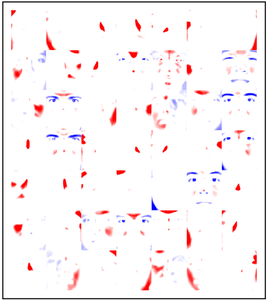
Faces: Extended Yale Database B



(d) SPCA, $\tau = 70\%$



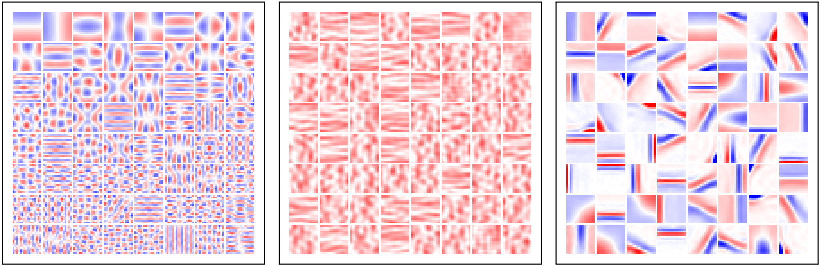
(e) SPCA, $\tau = 30\%$



(f) SPCA, $\tau = 10\%$

Extension to NMF and sparse PCA

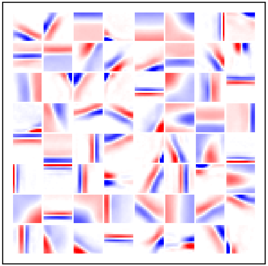
Natural Patches



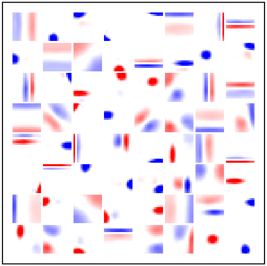
(a) PCA (b) NNMF (c) DL

Extension to NMF and sparse PCA

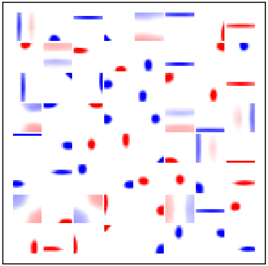
Natural Patches



(d) SPCA, $\tau = 70\%$



(e) SPCA, $\tau = 30\%$



(f) SPCA, $\tau = 10\%$

Conclusion

Take-home message

- Online techniques are adapted to the dictionary learning problem.
- Our method makes some large-scale image processing tasks tractable—. . .
- . . . — and extends to various matrix factorization problems.