Parsimonious Dictionary Learning

Mehrdad Yaghoobi, Thomas Blumensath, Mike E. Davies

Institute for Digital Communications, School of Engineering,
The University of Edinburgh, UK

ICASSP09, April 21, 2009

Outline

1. **Introduction**
   - Sparse Coding
   - Dictionary Selection Methods
   - Dictionary Learning for Sparse Approximations
2. **Parsimonious Dictionary Learning (PDL)**
   - Formulation
   - Algorithm
   - Update formula for $\mathbf{X}$
   - Update formula for $\mathbf{D}$
3. **Simulations**
   - Synthetic Data
   - PDL for Sparse Audio Coding
4. **Conclusion and Future Work**

Sparse Coding

Generative model

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \nu$$

$\mathbf{y} \in \mathbb{R}^d$, $\mathbf{D} \in \mathbb{R}^{d \times N}$, $\mathbf{x} \in \mathbb{R}^N$ and $\nu \in \mathbb{R}^d$. *Under-determined generative model*

$$\Leftrightarrow d < N$$

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} d_{1,1} & & d_{1,k} & & d_{1,N} \\ d_{2,1} & & d_{2,k} & & d_{2,N} \\ \vdots & \dots & \vdots & \dots & \vdots \\ d_{d,1} & & d_{d,k} & & d_{d,N} \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \\ \vdots \\ x_N \end{bmatrix}}_{\mathbf{x}} + \underbrace{\begin{bmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_d \end{bmatrix}}_{\nu}$$

- sparse coding: $\widehat{\mathbf{x}} = \arg\min_x \|\mathbf{x}\|_0 \text{ s. t. } \|\mathbf{y} - \mathbf{D}\mathbf{x}\|^2 \leq \xi$,
- $\xi = 0$ called sparse representation.
- $\xi > 0$ called sparse approximation.

Dictionary Selection Methods

- *Concatenation of orthonormal bases:* Let $\mathcal{O}$ be the set of all orthonormal dictionaries in $\mathbb{R}^{d \times d}$. $\mathcal{D} = \{\mathbf{D}_i\}_{i \in \mathcal{I}}, \forall i \in \mathcal{I}, \mathbf{D}_i \in \mathcal{O}$ is given. A dictionary $\mathbf{D}$ in $\mathbb{R}^{d \times d|\mathcal{I}|}$ is generated using,

$$\mathbf{D} = [\mathbf{D}_1 \cdots \mathbf{D}_i \cdots \mathbf{D}_{|\mathcal{I}|}].$$

- *Dictionary design subject to a certain property:* These properties include, but not restrict to, Restricted Isometry Property (RIP), minimum coherence μ and minimum cumulative coherence $\mu_1(m)$.

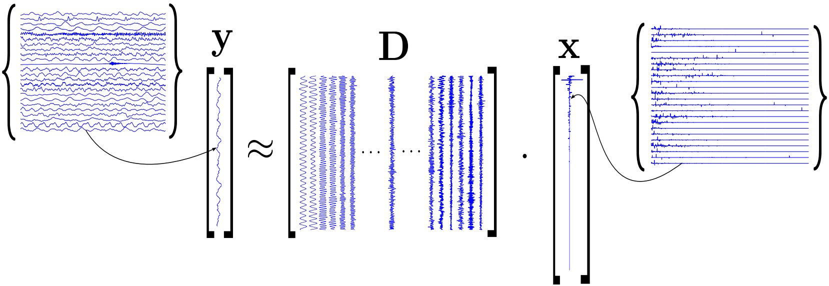
- *Dictionary learning using a set of training samples:* The goal is to find a dictionary such that it provides sparser coding for the given class of signals.

Dictionary Learning for Sparse Approximations

Definition

Let a set of training samples $\mathscr{L} = \{\mathbf{y}_i\}_{i \in \mathcal{I}}$ be given. Find a dictionary $\mathbf{D} \in \mathbb{R}^{d \times N}$ such that any training sample $\mathbf{y}_i$ has a sparse approximate representation $\mathbf{x}_i \in \mathbf{R}^N$ as follows,

$$\mathbf{y}_i \approx \mathbf{D}\mathbf{x}_i.$$



Dictionary Learning for Sparse Approximations as an optimization problem

Dictionary Learning for Sparse Approximations

The sparsity measure $\mathcal{J}(\mathbf{A}) = \sum_{i,j} |a_{i,j}|^{\rho},_{\rho \leq 1}$ and $\lambda \in \mathbb{R}^{+}$ is given.

$$\arg\min_{\mathbf{D}} \{\min_{\mathbf{X}} \phi(\mathbf{D}, \mathbf{X})\}$$

$$\phi(\mathbf{D}, \mathbf{X}) = \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_{\mathbf{F}}^{2} + \lambda \mathcal{J}(\mathbf{X})$$

Difficulties:

- *Scale Ambiguity:* $\forall (\alpha < 1) \in \mathbf{R}^{+}$, $\phi(\frac{1}{\alpha}\mathbf{D}, \alpha\mathbf{X}) \leq \phi(\mathbf{D}, \mathbf{X})$
- *Solution:* Constrained optimization, $\mathbf{D} \in \mathcal{D}$, where $\mathcal{D}$ is, for example, the constrained column or Frobenius norm dictionaries.
- *Model Order Ambiguity:* In model $\mathbf{Y}_{d \times L} \approx \mathbf{D}_{d \times N} \, \mathbf{X}_{N \times L}$, d and L are given and N is unknown in general.
- *Solution:* (our contribution) Applying a constraint on the dictionary size $\rightarrow$ learning a minimum size dictionary.

Parsimonious Dictionary Learning

Parsimonious Dictionary Learning: Formulation

$$\arg\min_{\mathbf{D}\in\mathcal{D}} \{\min_{\mathbf{X}} \phi(\mathbf{D}, \mathbf{X})\}$$

$$\phi(\mathbf{D}, \mathbf{X}) = \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_{\mathbf{F}}^2 + \lambda \mathcal{J}_{1,1}(\mathbf{X}) + \theta \mathcal{J}_{1,q}(\mathbf{D}^T)$$

Admissible Sets

- Bounded Frobenius-norm Dictionaries,

  $$\mathcal{D} = \{\mathbf{D}_{d\times N} : ||\mathbf{D}||_F \leq c_F^{1/2}\}$$

- Bounded Column-norm Dictionaries,

  $$\mathcal{D} = \{\mathbf{D}_{d\times N} : ||\mathbf{d}_i||_2 \leq c_C^{1/2}\}$$

Sparsity Measures

$$\mathcal{J}_{p,q}(\mathbf{A}) = \sum_{i\in I} [\sum_{j\in J} |a_{ij}|^q]^{\frac{p}{q}}$$

$$p \leq 1 \leq q$$

- $\mathcal{J}_{1,1}(\mathbf{A}) = \|\mathbf{A}\|_{\ell_1}$
- $\mathcal{J}_{1,2}(\mathbf{A})$: ℓ_1 norm of the ℓ_2 norms of the rows.

Parsimonious Dictionary Learning Algorithm

Parsimonious Dictionary Learning

$$\arg\min_{\mathbf{D}\in\mathcal{D}}\{\min_{\mathbf{X}}\phi(\mathbf{D},\mathbf{X})\}$$

$$\phi(\mathbf{D},\mathbf{X}) = \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_{\mathbf{F}}^2 + \lambda\mathcal{J}_{1,1}(\mathbf{X}) + \theta\mathcal{J}_{1,2}(\mathbf{D}^T)$$

- $\mathcal{D}$ convex set $\rightarrow \phi(\mathbf{D},\mathbf{X})$ is bi-convex, i.e. convex w.r.t each parameter, when the other is kept fixed.
- $\phi(\mathbf{D},\mathbf{X})$ can be minimized using alternating minimization technique.
- Optimization w.r.t each parameter can be done using convex optimization methods $\rightarrow$ Majorization Minimization Method.
- The quadratic term $\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2$ couples the components of $\mathbf{D}$ and $\mathbf{X}$ such that the element-wise optimization of $\phi(\mathbf{D},\mathbf{X})$ becomes difficult. Majorization minimization simplifies the optimization by de-coupling the quadratic term.

Majorization Method

Majorization minimization method: replacing the original objective $\phi(\omega)$ with the surrogate majorizing objective $\psi(\omega, \xi)$.

Optimization problem

$$\min_{\omega \in \Omega} \phi(\omega)$$

$$c \leq \phi(\omega)$$

Majorizing objective

$$\phi(\omega) \leq \psi(\omega, \xi) \quad \forall \omega, \xi \in \Omega$$

$$\phi(\omega) = \psi(\omega, \omega) \quad \forall \omega \in \Omega$$

Two-step optimization

1- $\omega_{new} = \arg\min_{\omega \in \Omega} \psi(\omega, \xi)$, *fixed* ξ

2- $\xi_{new} = \omega = \arg\min_{\xi \in \Omega} \psi(\omega, \xi)$, *fixed* ω

- The surrogate objective can be found by adding a strictly convex function, with a minimum at $\omega = \xi$, to the original objective.

Update formula for X

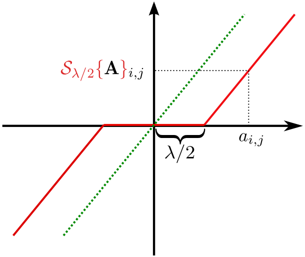
- Let $\phi_{\mathbf{D}}(\mathbf{X})$ be the objective while $\mathbf{D}$ is kept fixed.
- The function $\pi_{\mathbf{X}}(\mathbf{X}, \mathbf{X}^{[n]}) := c_X||\mathbf{X} - \mathbf{X}^{[n]}||_F^2 - ||\mathbf{DX} - \mathbf{DX}^{[n]}||_F^2$, which is strictly convex for a selected c_X and has a minimum at $\mathbf{X} = \mathbf{X}^{[n]}$, is added to $\phi_{\mathbf{D}}(\mathbf{X})$ to generate the surrogate objective $\psi_{\mathbf{D}}(\mathbf{X}, \mathbf{X}^{[n]})$.
- $\psi_{\mathbf{D}}(\mathbf{X}, \mathbf{X}^{[n]})$ is convex w.r.t $\mathbf{X}$ and $\mathbf{0}$ is in the subgradient at the minimum.

$$\mathbf{0} \in \partial\psi_{\mathbf{D}}(\mathbf{X}^{[n+1]}, \mathbf{X}^{[n]}),$$

$$\partial\psi_{\mathbf{D}}(\mathbf{X}, \mathbf{X}^{[n]}) = 2c_{\mathbf{X}}\mathbf{X} - 2(\mathbf{D}^T(\mathbf{Y} - \mathbf{DX}^{[n]}) + c_{\mathbf{X}}\mathbf{X}^{[n]}) + \lambda\partial\mathcal{J}_{1,1}(\mathbf{X}),$$

$$\therefore \mathbf{X}^{[n+1]} = \mathcal{S}_{\lambda/2}\{\mathbf{A}\}$$

$$\mathbf{A} = \frac{1}{c_{\mathbf{X}}}(\mathbf{D}^T(\mathbf{Y} - \mathbf{DX}^{[n]}) + c_{\mathbf{X}}\mathbf{X}^{[n]})$$



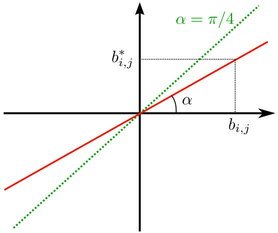
Update formula for D

- Let $\phi_{\mathbf{x}}(\mathbf{D})$ be the objective while $\mathbf{X}$ is kept fixed.
- The surrogate objective:

$$\psi_X(\mathbf{D}, \mathbf{D}^{[n]}) = \phi_{\mathbf{x}}(\mathbf{D}) + \pi_{\mathbf{D}}(\mathbf{D}, \mathbf{D}^{[n]}),$$
$$\pi_{\mathbf{D}}(\mathbf{D}, \mathbf{D}^{[n]}) := c_D||\mathbf{D} - \mathbf{D}^{[n]}||_F^2 - ||\mathbf{D}\mathbf{X} - \mathbf{D}^{[n]}\mathbf{X}||_F^2$$

- $\psi_{\mathbf{x}}(\mathbf{D}, \mathbf{D}^{[n]})$ is convex w.r.t $\mathbf{D}$ and $\mathbf{0}$ is in the subgradient at the minimum.

$$\mathbf{0} \in \partial\psi_{\mathbf{x}}(\mathbf{D}^{[n+1]}, \mathbf{D}^{[n]}),$$
$$\partial\psi_{\mathbf{x}}(\mathbf{D}, \mathbf{D}^{[n]}) = 2c_D\mathbf{D} - 2((\mathbf{Y} - \mathbf{D}^{[n]}\mathbf{X})\mathbf{X}^T + c_D\mathbf{D}^{[n]}) + \theta\ \partial\mathcal{J}_{1,2}(\mathbf{D}^T)$$
$$\therefore \mathbf{D}^{[n+1]} = \mathcal{P}_{\mathcal{D}}\{\mathbf{B}^*\}$$
$$\mathbf{B}^* = \mathcal{O}_{\frac{\theta}{c_D}}\{\mathbf{B}\}$$
$$\mathbf{B} = \frac{1}{c_D}((\mathbf{Y} - \mathbf{D}^{[n]}\mathbf{X})\mathbf{X}^T + c_D\mathbf{D}^{[n]})$$

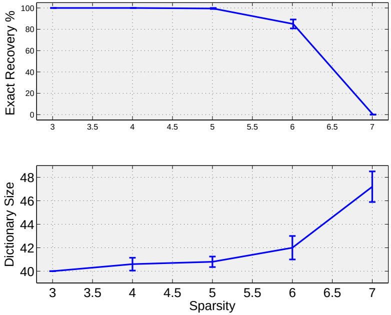


$$\tan(\alpha) = \begin{cases} 1 - \frac{\theta}{2c_D||\mathbf{b}_j||_2} & \frac{\theta}{2c_D} < ||\mathbf{b}_j||_2 \\ 0 & \textit{otherwise} \end{cases}$$

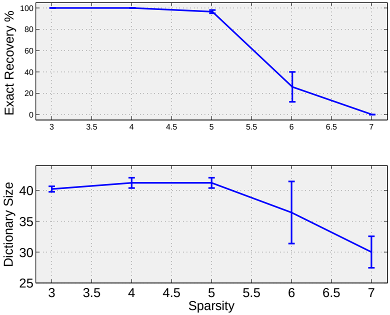
Simulations: Exact Dictionary Recovery

d	N	L	λ	θ	T	$\mathbf{D}^{[0]}$	N_0	$\mathbf{X}^{[0]}$
20	40	1280	0.4	0.05	1000	$\mathcal{N}(0,1)$	80	$\mathbf{0}$

$\mathcal{D} = \{\mathbf{D}_{d\times N} : ||\mathbf{d}_i||_2 \leq 1\}$

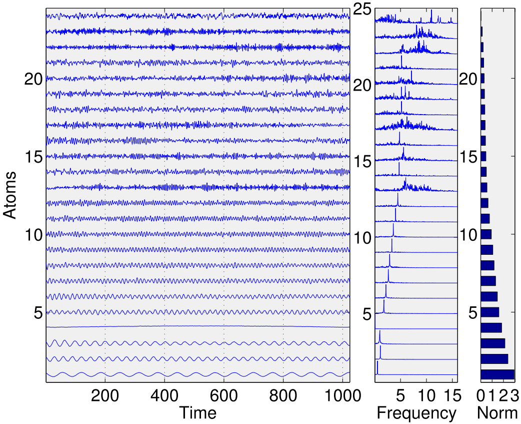


$\mathcal{D} = \{\mathbf{D}_{d\times N} : ||\mathbf{D}||_F \leq \sqrt{N}\}$



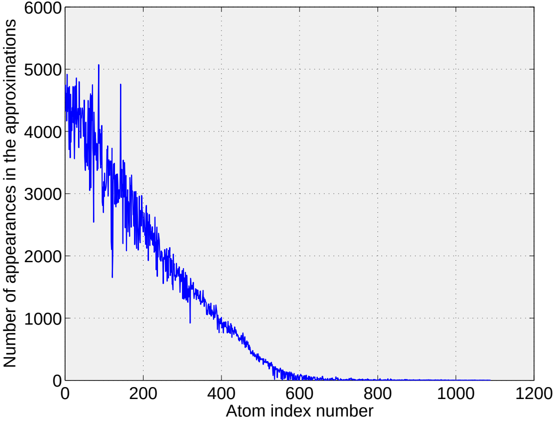
Dictionary Learning for Audio Coding

d	N	L	$\mathcal{D}$	T	$\mathbf{D}^{[0]}$	$\mathbf{X}^{[0]}$
1024	2048	8192	$\lVert\mathbf{D}\rVert_F \leq \sqrt{N}$	5000	$2 \times DCT$	$\mathbf{0}$



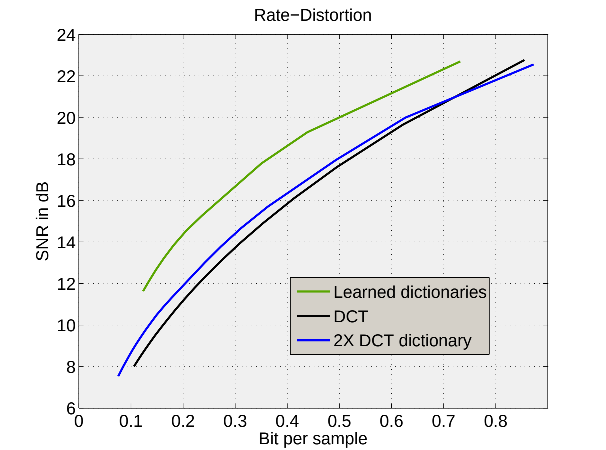
$\lambda = 0.01 \qquad \theta = 0.01$

Number of Appearances of Learned Atoms in the Approximations



$\boxed{L = 8192 \qquad \lambda = 0.01}$

Rate-Distortion of the Audio Coding using Different Dictionaries



$$\boxed{L = 4096 \qquad N : \textit{varies}}$$

Conclusion and Future Work

Conclusion

- A new framework for dictionary learning, under a minimum order constraint, was presented.
- A practical algorithm was presented to *approximately* solve the non-convex optimization problem.
- By some simulations, on the synthetic data, it has been shown that the algorithm recovers correct atoms and correct dictionary size .
- The learned dictionary, using samples of audio signals, has shown a superior performance in the sparse audio coding, in terms of Rate-Distortion.

Future Work

- Finding an automatic method to adjust θ.
- Extending the framework to a parsimonious dictionary selection.
- Using an alternative, and non-convex, sparsity measure.

Thanks for your attention.

Any questions?